Python *for* Econometrics

Lecturer: Fabian H. C. Raters

Institute: Econometrics, University of Goettingen

Version: October 08, 2018

Welcome to this course and to the world of Python!

Learning objectives of this course:

- Python: Roughly half the course is about Python.
- *for*: You will learn tools and methods.
- Econometrics:
  - Statistics: Numerical programming in Python.
  - applied to: We will use it on examples.
  - Economics: In an economic context.

Knowledge after completing this course:

- You have acquired a basic understanding of programming in general with Python and a special knowledge of working with standard numerical packages.
- You are able to study Python in depth and absorb new knowledge for your scientific work with Python.
- You know the capabilities and further possibilities to use Python in econometrics.

What you should not expect from this course:

- A guide how to install or maintain an application.
- An introduction to programming for beginners.
- Non-scientific, general purpose programming (beyond the language essentials).
- Introduction to professional development tools.
- Few content and less effort...

This course can be seen as an applied lecture:

Lecture:
We try to explain the partly theoretical knowledge on Python by simple, easy to understand examples. You can learn the subtleties by reading good literature.

Exercises:
Digital work sheets in the form of Jupyter notebooks with applied tasks are available for each chapter. For all exercises there are sample solutions available in separate notebooks.

Self-tests:
At the end of each of the five chapters there are typical exam questions.

Written exam:
There will be a final exam. This will be a pure multiple choice exam: 60 questions, 90 minutes.

After the successful participation in the exam you will receive 6 ECTS.

Literature

The programming language Python is already established and very well in trend for numerical applications. Some keywords:

- Data science,
- Data wrangling,
- Machine learning,
- Numerical statistics,
- ...

Recommended literature while following this course:

- *Learning Python, 5th Edition* by Mark Lutz,
- *Python Crash Course* by Eric Matthes,
- *Python Data Science Handbook* by Jake VanderPlas,
- *Python for Data Analysis, 2nd Edition* by Wes McKinney,
- *Python for Finance* by Yves Hilpisch.

We are using *Python 3*. There was a big revision in the migration from Python 2 to version 3 and the new version is no longer backwards compatible to the old version.

Python 3 running [command line]

```
python3 --version


## Python 3.6.6
```

The normal execution mode is that the Python interpreter processes the instructions in the background – in other numeric programming languages such as *R* this is known as *batch mode*. It executes program code that is usually located in a source code file.

The interpreter can also be started in an *interactive mode*. It is used for testing and analytical purposes in order to obtain fast results when performing simple applications.

For everyday work with Python it would be extremely tedious to make all edits in interactive mode.

There are a number of excellent integrated development environments (IDEs) for Python, with two being emphasized here:

- *Jupyter* (and *IPython*)
- *PyCharm* (by *IntelliJ*)

Of course, you can also use a simple text editor. However, you would probably miss the comfort of an IDE.

Installing, adding and maintaining Python is not trivial at the beginning. Therefore, as a beginner, you are well advised to download and install the Python distribution *Anaconda*. Bonus: Many standard packages are supplied directly or you can post-install them conveniently.

In this course – in a numerical and analytical context – we use only Jupyter with the IPython kernel.

That is why we have combined

1. all the code from the slides, and
2. all the exercises and solutions

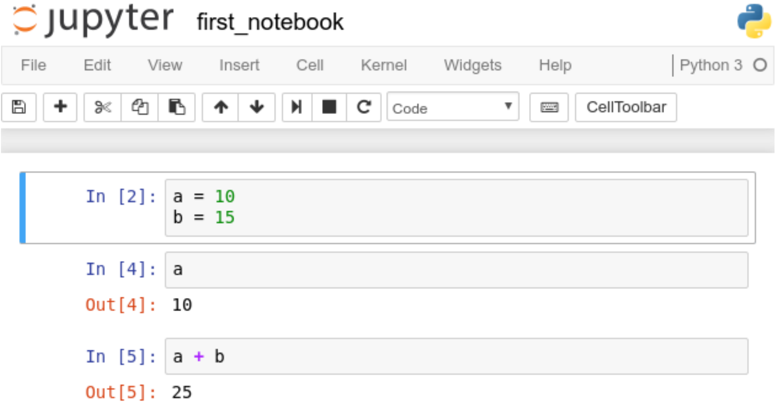
into interactive Jupyter notebooks that you can use online without having to install software locally on your computer. The GWDG has set up a cloud-based *Jupyter-Hub* for you.

You can access the working environment with your university credentials at

https://jupyter.gwdg.de/

create a profile and get started right away – even using your smart devices. However, so far you are still asked to upload the course notebooks by yourself or rewrite the code from scratch.

A Jupyter notebook is divided into individual, vertically arranged cells, which can be executed separately:



The notebook approach is not novel and comes from the field of computer algebra software.

Actually, an interactive Python interpreter called IPython is started "in the core".

IPython running [command line]

```
ipython3 --version
```

```
## 6.5.0
```

Roughly speaking, this is a greatly enhanced version of the Python 3 interpreter, which has numerous, convenient advantages over the "normal" interpreter in interactive mode, such as, e. g.,

- printing of return values,
- color highlighting, and
- magic commands.

Finally, we wish you a lot of fun and success with and in this course!

Practice makes perfect!

Contribution and credits:

Fabian H. C. Raters

Eike Manßen

GWDG *for the Jupyter-Hub*

Essential concepts

Essential concepts

▶ Getting started

Python can be described as

- a dynamic, strongly typed, multi-paradigm and object-oriented programming language,
- for versatile, powerful, elegant and clear programming,
- with a general, high-level, multi-platform application scope,
- which is being used very successfully in the data science sector and very much in trend.

Moreover, Python is relatively easy to learn and its successful language design supports novices to professional developers.

... of the Python era:

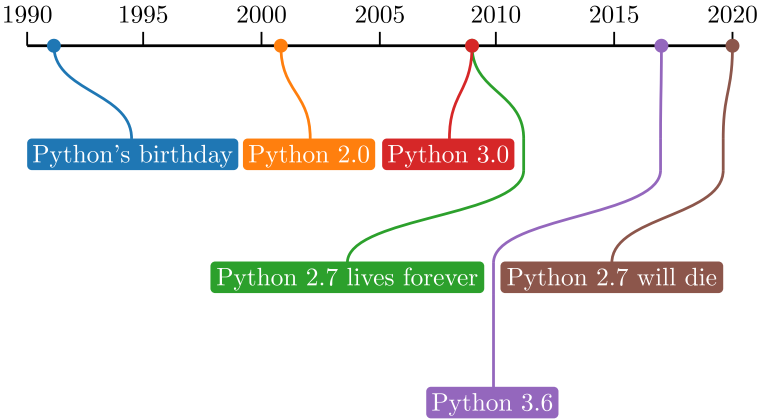
The language was originally developed in 1991 by Guido van Rossum. Its name was based on Monty Python's Flying Circus. Its main identification feature is the novel markup of code blocks – by indentation:

Indentation example

```
password = input("I am your bank. Password please: ")
## I am your bank. Password please: sparkasse

if password == "sparkasse":
    print("You successfully logged in!")
else:
    print("Fail. Will call the police!")

## You successfully logged in!
```

This increases the readability of code and should at the same time encourage the programmer in programming neatly. Since the source code can be written more compactly with Python, an increased efficiency in daily work can be expected.

Overview of the Python development by versions and dates:



In comparison

Comparing the way Python works with common programming languages, we briefly discuss a selection of popular competitors:

C/C++:

- CPython is interpreted, not compiled.
- C/C++ are strongly static, complex languages.

Java:

- CPython is not compiled just-in-time.
- Java has a *C*-type syntax.

MATLAB

- In Python you primarily follow a scalar way of thinking, while in *MATLAB* you write matrix-based programs.
- In the numerical context, the matrix view and syntax are very similar to those of MATLAB.
- MATLAB is partially compiled just-in-time.

Where *CPython* is the reference implementation – the "Original Python", which is implemented in C itself.

R

- In Python you primarily follow a scalar way of thinking, while in *R* you write vector-based programs.
- R has a C-type syntax including additions to novel language concepts.

Stata

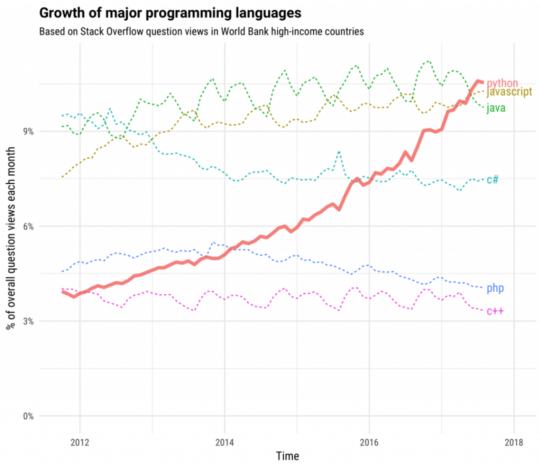
- Any comparison would inadequately describe the differences.

Reference semantics

An extremely important difference between the first two languages, C/C++ and Java, as well as Python itself, and the last three languages is that they follow a call-by-reference semantic, while MATLAB, R and Stata are call-by-copy.

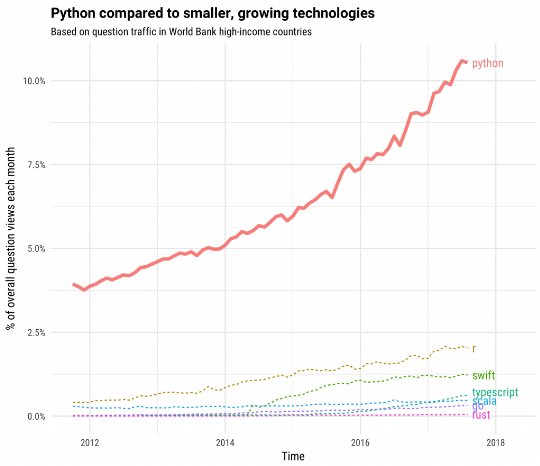
Further specific differences and similarities to MATLAB and R will be addressed in other parts of this course.

Python has become extremely popular:



Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

So, you're on the right track – because who wants to bet on the wrong ho*R*se?



Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Versatility – diversity

Areas in which Python is used with great success:

- Scripts,
- Console applications,
- GUI applications,
- Game development,
- Website development, and
- Numerical programming.

Places where Python is used:

In this course we will successively gain the following insights:

1. General basics of the language.
2. Numerical programming and handling of data sets.
3. Application to economic and analytical questions.

Essential concepts

- Procedural programming

Programs can be implemented very quickly – this is a pretty minimal example. You can write this command to a text file of your choice and run it directly on your system:

Hello there

```
print("Hello there!")

## Hello there!
```

- Only one *function* `print()` (shown here as a *keyword*),
- Function displays *argument* (a string) on screen,
- Arguments are passed to the function in parentheses,
- A string must be wrapped in `" "` or `' '`,
- No semicolon at the end.

User input

Let's add a user input to the program:

Hello you

```
name = input("Please enter your name: ")
## Please enter your name: Angela Merkel
print("Hello " + name + "!")

## Hello Angela Merkel!
```

- The function `input()` is used for interactive text input,
- You can use the equal sign = to assign variables (here: `name`),
- Strings can be joined by the (overloaded) Operator +.

We are now trying to find out on which weekday a person was born (Merkel's birthday is 17-07-1954):

Weekday of birth

```
from datetime import datetime

answer = input("Your birthday (DD-MM-YYYY): ")

## Your birthday (DD-MM-YYYY): 17-07-1954

birthday = datetime.strptime(answer, "%d-%m-%Y")
print("Your birthday was on a " + birthday.strftime("%A") + "!")

## Your birthday was on a Saturday!
```

- It is really easy to import functionality from other *modules*,
- Function `strptime()` is a *method* of *class* `datetime`,
- Both methods, `strptime()` and `strftime()`, are used to convert between strings and date time specifications.

And how many days have passed since then (until Merkel's 4th swearing-in as Federal Chancellor)?

Age in days

```
someday = datetime.strptime("09-10-2018", "%d-%m-%Y")
print("You are " + str((someday - birthday).days) + " days old!")

## You are 23460 days old!
```

- You can create time differences, i. e. the operator − is overloaded,
- The difference represents a new *object*, with its own *attributes*, such as `days`,
- When using the overloaded operator +, you have to explicitly convert the number of days by means of `str()` into a string.

How many years, weeks and days do you think that is?

Human readable age

```
from dateutil.relativedelta import relativedelta
delta = relativedelta(someday, birthday)
print(f"That’s {delta.years} years, {delta.months} months "
      f"and {delta.days} days!!")

## That's 64 years, 2 months and 22 days!!
```

- You don’t have to keep reinventing the wheel – a wealth of packages and individual modules are freely available,
- A lowercase `f` before `"..."` provides convenient *formatting* – there are other options as well,
- Two strings in sequence are implicitly joined together – `"That" "’s nice"`!

When working with the interactive interpreter, i. e. in a notebook, you can quickly get useful information about Python objects:

Help system

```
help(len)

## Help on built-in function len in module builtins:
##
## len(obj, /)
##     Return the number of items in a container.
```

Alternatively, e. g., for more complex problems, it is best to search directly with your preferred internet search engine.

You can find neat solutions to conventional challenges in literature.

As with natural language, programming languages have a lexical structure. Source code consists of the smallest possible, indivisible elements, the tokens. In Python you can find the following groups of elements:

- Literals
- Variables
- Operators
- Delimiters
- Keywords
- Comments

These terms give us a rock-solid foundation for exploring the heart of a programming language.

Basically, we distinguish between *literals* and *variables*:

Assigning variables with literals

```
myint = 7
myfloat = 4.0
myboat = "nice"
mybool = True
myfloat = myboat
```

- In this course, we will work with four different literals: integer (`7`), float (`4.0`), string (`"nice"`) and boolean (`True`),
- Literals are assigned to variables at runtime,
- In Python the data type is derived from the literal and does not have to be described explicitly,
- It is allowed to assign values of different data types to the same variable (name) sequentially,
- If we don't assign a literal to any variables, we forfeit it.

Operators and delimiters

Most *operators* and *delimiters* will be introduced to you during this course. Here is an overview of the operators:

Overview of operators

```
## +       -       *       /       **      //
## %       <<      >>      &       |       ^
## ~       and     or      not     in      not in
## is      is not  <       >       !=      <>
## ==      <=      >=
```

An overview of the delimiters follows:

Overview of delimiters

```
## (       )       [       ]       {       }
## ,       :       .       `       =       ;
## +=      -=      *=      /=      **=     //=
## %=      <=      |=      ^=      >>=     <<=
## '       "       \       @       SPACE
```

All regular *arithmetic operations* involving numbers are possible:

Pocket calculator

```
10 + 5
100 - 20
8 / 2
4 * (10 + 20)
2**3
```

```
## 15
## 80
## 4.0
## 120
## 8
```

- The result of dividing two integers is a floating point number,
- The conventional rules apply: Parentheses first, then multiplication and division, etc.,
- The operator `**` is used for exponentiation.

The programmer explains the structure of his/her program to the interpreter via a restricted set of short commands, the *keywords*:

Overview of keywords

```
## and       as    assert  break   class     continue
## def       del   elif    else    except    False
## finally   for   from    global  if        import
## in        is    lambda  None    nonlocal  not
## or        pass  raise   return  True      try
## while     with  yield
```

There are two ways to make *comments*:

Give some comments

```
# Set variable to something - or nothing?
something = None
"""
I am a docstring!
A multiline string comment hybrid.
I will be useful for describing classes and methods.
"""
```

We can create a handy table summarizing some results demonstrating the use of *logical operators* (and formatted strings and `for`-loops):

Logical table

```
# Create table head
print("a    b    a and b    a or b    not a\n"
      "-----------------------------------")
# Loop through the rows
for a in [False, True]:
    for b in [False, True]:
        print(f"{a:1} {b:3} {a and b:6} {a or b:8} {not a:7}")
```

```
## a    b    a and b    a or b    not a
## -----------------------------------
## 0    0       0          0        1
## 0    1       0          1        1
## 1    0       0          1        0
## 1    1       1          1        0
```

Python offers the following *basic data types*, which we will use in this course:

Data type	Description
`int()`	Integers
`float()`	Floating point numbers
`str()`	Strings, i.e. unicode (UTF-8) texts
`bool()`	Boolean, i.e. `True` or `False`
`list()`	List, an ordered array of objects
`tuple()`	Tuple, an ordered, unmutable array of objects
`dict()`	Dictionary, an unordered, associative array of objects
`set()`	Set, an unordered array/set of objects
`None()`	Nothing, emptyness, the void..

Each data type has its own methods, that is, functions that are applicable specifically to an object of this type.

You will gradually get to know new and more complex data types or object classes.

A *list* is an ordered array of objects, accessible via an *index*:

Listing tech companies

```
stocks = ["Google", "Amazon", "Facebook", "Apple"]
stocks[1]
stocks.append("Twitter")
stocks.insert(2, "Microsoft")
stocks.sort()
```

```
## ['Google', 'Amazon', 'Facebook', 'Apple']
## Amazon
## ['Google', 'Amazon', 'Facebook', 'Apple', 'Twitter']
## ['Google', 'Amazon', 'Microsoft', 'Facebook', 'Apple', 'Twitter']
## ['Amazon', 'Apple', 'Facebook', 'Google', 'Microsoft', 'Twitter']
```

- The constructor for new lists is `[ ]`,
- The first element has the index 0,
- The data type `list()` possesses its own methods.

Tuples

Tuples are immutable sequences related to lists that cannot be extended, for example. The drawbacks in flexibility are compensated by the advantages in speed and memory usage:

Selecting elements in sequences

```
lottery = (1, 8, 9, 12, 24, 28)
len(lottery)
lottery[1:3]
lottery[:4]
lottery[-1]
lottery[-2:]
```

```
## (1, 8, 9, 12, 24, 28)
## 6
## (8, 9)
## (1, 8, 9, 12)
## 28
## (24, 28)
```

The same operations are also supported when using lists.

Dictionaries are associative collections of *key-value pairs*. The *key* must be immutable and unique:

Internet slang dictionary

```
slang = {"imho": "in my humble opinion",
         "lol": "laughing out loud",
         "tl;dr": "too long; didn’t read"}
slang["lol"]
slang["gl&hl"] = "good luck & have fun"
slang.keys()
slang.values()
```

```
## {'imho': 'in...ion', 'lol': 'la...oud', 'tl;dr': 'to...ead'}
## laughing out loud
## good luck & have fun
## dict_keys(['imho', 'lol', 'tl;dr', 'gl&hl'])
## dict_values([... & have fun'])
```

- The constructor for `dict()` is `{ }` with `:`,
- The pairs are unordered, iterable sequences.

A *set* is an unordered collection of objects without duplicates:

Set operations

```
x = {"o", "n", "y", "t"}
y = {"p", "h", "o", "n"}
x & y
x | y
x - y
```

```
## {'y', 'o', 'n', 't'}
## {'h', 'o', 'p', 'n'}
## {'o', 'n'}
## {'y', 'n', 'h', 'o', 'p', 't'}
## {'y', 't'}
```

- The constructor for `set()` is `{ }`,
- Defines its own operators that overload existing ones.
- Empty set via `set()`, because `{}` already creates `dict()`.

Python has only one kind of conditional statement – `if-elif-else`:

Computer data sizes

```
bytes = 100000000 / 8  # e.g. DSL 100000
if bytes >= 1e9:
    print(f"{bytes/1e9:6.2f} GByte")
elif bytes >= 1e6:
    print(f"{bytes/1e6:6.2f} MByte")
elif bytes >= 1e3:
    print(f"{bytes/1e3:6.2f} KByte")
else:
    print(f"{bytes:6.2f} Byte")

##  12.50 MByte
```

Control flow structures may be nested in any order:

Nestings

```
if a > 1:
    if b > 2:
        pass  # special keyword for empty blocks
```

In Python there exist two conventional *program loops* – `for-in-else`:

Total sum

```
numbers = [7, 3, 4, 5, 6, 15]
y = 0
for i in numbers:
    y += i
print(f"The sum of ’numbers’ is {y}.")

## The sum of 'numbers' is 40.
```

Lists or other collections can also be created dynamically:

Powers of 2

```
powers = [2 ** i for i in range(11)]
teacher = ["***", "**", "*"]
grades = {star: len(teacher) - len(star) + 1 for star in teacher}

## [1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]
## {'***': 1, '**': 2, '*': 3}
```

Loops can skip iterations (`continue`):

Continue the loop

```
for x in ["a", "b", "c"]:
    a = x.upper()
    continue
    print(x)
print(a)

## C
```

Or a loop can be aborted instantly (`break`):

Breaking the habit

```
y = 0
for i in [7, 3, 4, "x", 6, 15]:
    if not isinstance(i, int):
        break
    y += i
print(f"The total sum is {y}.")

## The total sum is 14.
```

Control flow: while loop

For loops where the number of iterations is not known at the beginning, you use `while`-`else`.

Have you already noticed the keyword `else`? Python only executes the branch if it was not terminated by `break`:

Favorite lottery number

```
import random
n = 0
favorite = 7
while n < 100:
    n += 1
    draw = random.randint(1, 49)  # e.g. German lottery
    if draw == favorite:
        print("Got my number! :)")
        break
else:
    print("My favorite did not show up! :(")
print(f"I tried {n} times!")
```

```
## Got my number! :)
## I tried 15 times!
```

Functions are defined using the keyword `def`. The structure of *function signature* and *body* is specified by indentation, too:

Drawing lottery numbers

```
def draw_sample(n, first=1, last=49):
    numbers = list(range(first, last + 1))
    sample = []
    for i in range(n):
        ind = random.randint(0, len(numbers) - 1)
        sample.append(numbers.pop(ind))
    sample.sort()
    return sample

draw_sample(6)
draw_sample(6, 80, 100)
draw_sample(3, first=5)

## [26, 31, 33, 36, 41, 49]
## [83, 84, 85, 87, 91, 92]
## [18, 25, 42]
```

Functions are of type `callable()`, defined as closures, and can be created and used like other objects:

Prime numbers

```python
def primes(n):
    numbers = [2]
    def is_prime(num):
        for i in numbers:
            if num % i == 0:
                return False
        return True
    if n == 2:
        return numbers
    for i in range(3, n + 1):
        if is_prime(i):
            numbers.append(i)
    return numbers
```

```python
primes(50)
```

```
## [2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47]
```

Seems weird? We discuss namespaces in the next section.

Essential concepts

- Object-orientation

There are three widely known programming paradigms: *procedural*, *functional* and *object-oriented programming* (*OOP*). Python supports them all.

You have learned how to handle predefined data types in Python. Actually, we have already encountered classes and instances, take for example `dict()`.

In this section you will learn the basics of dealing with (your own) classes:

1. References
2. Classes
3. Instances
4. Main principles
5. Garbage collection

OOP is a wide field and challenging for beginners. Don't get discouraged and, if you find deficits in yourself, read the literature.

When you assign a variable, a *reference* to an object is set:

Equal but not identical

```
a = ["Star", "Trek"]
b = ["Star", "Trek"]
c = a
a == b
a == c
a is b
a is c
```

```
## ['Star', 'Trek']
## ['Star', 'Trek']
## ['Star', 'Trek']
## True
## True
## False
## True
```

- Two equal but not identical objects are created,
- Variables `a` and `c` link to the same object.

When we introduced lists, we initially did not mention that they are a first-class example of *mutable* objects:

Collecting grades

```
grades = [1.7, 1.3, 2.7, 2.0]
result = grades.append(1.0)
result
grades
finals = grades
finals.remove(2.7)
finals
grades
```

```
## None
## [1.7, 1.3, 2.7, 2.0, 1.0]
## [1.7, 1.3, 2.0, 1.0]
## [1.7, 1.3, 2.0, 1.0]
```

- Modifications can be *in-place* – the object itself is modified.
- Changing an object that is referenced several times could cause (un)intended consequences.

In Python, arguments are *passed by assignment*, i.e. call-by-reference:

Side effects

```
def last_element(x):
    return x.pop(0)
a = stocks
last_element(a)
a
```

```
## ['Amazon', 'Apple', 'Facebook', 'Google', 'Microsoft', 'Twitter']
## Twitter
## ['Amazon', 'Apple', 'Facebook', 'Google', 'Microsoft']
```

- There are side effects,
- Referenced *mutable* objects might be modified,
- Referenced *immutable* objects might be copied.

We are able to make an exact copy of the object:

Copying

```
def last_element(x):
    y = x.copy()
    return y.pop(-1)
a = stocks
last_element(a)
a

## ['Amazon', 'Apple', 'Facebook', 'Google', 'Microsoft']
## Microsoft
## ['Amazon', 'Apple', 'Facebook', 'Google', 'Microsoft']
```

- We receive a new object,
- The new object is not identical to the old one.

However, keep in mind that, in most cases, a method `copy()` will create *shallow* copys while only *deep copying* will duplicate also the contents of a mutable object with a complex structure:

Cloning fast food

```
fastfood = [["burgers", "hot dogs"], ["pizza", "pasta"]]
italian = fastfood.copy()
italian.pop(0)
american = list(fastfood)
american.pop(1)
american[0] = american[0].copy()
fastfood[0][1] = "chicken wings"
fastfood[1][0] = "risotto"
italian
american

## [['risotto', 'pasta']]
## [['burgers', 'hot dogs']]
```

Both approaches, `copy()` and `list()`, create new `list` objects containing new references to the original sub-lists. But for a *deep copy*, you have to recursively create duplicates of all its objects.

Classes

In Python everything is an object and more complex objects consist of several other objects.

In the OOP, we create objects according to patterns. These kinds of blueprints are called *classes* and are characterized by two categories of elements:

Attributes:
Variables that represent the properties of

- an object, *object attributes*, or
- a class, named *class attributes*.

Methods:
Functions that are defined within a class:

- *(non-static) methods* can access all attributes, while
- *static methods* can only access class attributes.

Every generated object is an *instance* of such a construction plan.

Specifically, we want to create "rectangle object" and define a separate `Rectangle` class for it:

Rectangle class

```
class Rectangle:
    width = 0
    height = 0
    def area(self):
        return self.width * self.height
myrectangle = Rectangle()
myrectangle.width = 10
myrectangle.height = 20
print(myrectangle.area())

## 200
```

- New classes are defined using the keyword `class`,
- The variable `self` always refers to the instance itself.

We add a *constructor* (method) `__init__()`, that is called to *initialize* an object of `Rectangle`:

Rectangle class with constructor

```
class Rectangle:
    width = 0
    height = 0
    def __init__(self, width, height):
        self.width = width
        self.height = height
    def area(self):
        return self.width * self.height
myrectangle = Rectangle(15, 30)
print(myrectangle.area())

## 450
```

In our example, we use the constructor to set the attributes. Methods with names matching `__fun__()` have a special, standardized meaning in Python.

One of the most important concepts of OOP is *inheritance*. A class inherits all attributes and methods of its *parent class* and can *add new* or *overwrite* existing ones:

Square inherits Rectangle

```
class Square(Rectangle):
    def __init__(self, length):
        super().__init__(length, length)
    def diagonal(self):
        return (self.width**2 + self.height**2)**0.5
mysquare = Square(15)

print(f"Area: {mysquare.area()}")
print(f"Diagonal length: {mysquare.diagonal():7.4f}")

## Area: 225
## Diagonal length: 21.2132
```

The methods of the parent class, including the constructor, may be referenced by `super()`.

You do not have to worry about memory management in Python. The *garbage collector* will tidy up for you.

If there are no more references to an object, it is automatically disposed of by the garbage collector:

Garbage collection in action

```
class Dog:
    def __del__(self):
        print("Woof! The dogcatcher got me! Entering the void.. :(")
# My old dog on a leash
mydog = Dog()
# A new dog is born
newdog = Dog()
# Using my leash for the new dog
mydog = newdog

## Woof! The dogcatcher got me! Entering the void.. :(
```

The *destructor* `__del__()` is executed as the last act before an object gets deleted.

Namespaces

We have already come into contact with *namenspaces* in Python many times. These are hierarchically linked layers in which the references to objects are defined. A rough distinction is made between

- the *global* namespace, and
- the *local* namespace.

The global namespace is the *outermost environment* whose references are known by all objects.

On the other hand, locally defined references are only known in a local, i.e. *internal environment*.

Reference names from the local namespace mask the same names in an outer or in the global namespace:

Namespaces

```
def multiplier(x):
    x = 4 * x
    return x
x = "OH"
multiplier("AH")
multiplier(x)
x
```

```
## OH
## AHAHAHAH
## OHOHOHOH
## OH
```

In fact, functions defined in Python are themselves objects that remember and can access their own context where they were created. This concept comes from functional programming and is called *closure*:

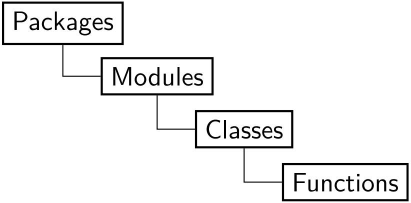
Closures

```
def gen_multiplier(a):
    def fun(x):
        return a * x
    return(fun)
multi1 = gen_multiplier(4)
multi2 = gen_multiplier(5)
multi1
multi1("EH")
multi2("EH")
```

```
## <function gen_multiplier.<locals>.fun at 0x7f042eaa6048>
## EHEHEHEH
## EHEHEHEHEH
```

Managing code

In order to provide, maintain and extend modular functionality with Python, its code containing components can be described hierarchically:



The organization in Python is very straightforward and is based on the local namespaces mentioned before.

When you download and use new *packages*, such as *NumPy* for numerical programming in the next chapter, the packages are loaded and the namespaces initialized.

The development of custom packages is an advanced topic and not essential for a reasonable code structure of small projects, as it is in other programming languages.

Modules provide classes and functions via namespaces. It is Python code that is executed in a local namespace and whose classes and functions you can import. Basically, there are the following alternatives how to *import* from an module:

Import statements

```
import datetime as dt
from datetime import date, timedelta
from datetime import *
dt.date.today()
dt.timedelta.days
date.today()
timedelta.days
datetime.now()
```

In the latter case, all classes and functions, but no instances, are imported from the `datetime` namespace.

The Zen of Python

```
import this
```

```
## The Zen of Python, by Tim Peters
##
## Beautiful is better than ugly.
## Explicit is better than implicit.
## Simple is better than complex.
## Complex is better than complicated.
## Flat is better than nested.
## Sparse is better than dense.
## Readability counts.
## Special cases aren't special enough to break the rules.
## Although practicality beats purity.
## Errors should never pass silently.
## Unless explicitly silenced.
## In the face of ambiguity, refuse the temptation to guess.
## ...
```

A selection of exciting topics that are among the advanced basics but are not covered in this lecture:

- Dynamic language concepts, such as duck typing,
- Further, complex type classes, such as `ChainMap` or `OrderedDict`,
- Iterators and generators in detail,
- Exception handling, raising exceptions, catching errors,
- Debugging, introspection and annotations.

Numerical programming

2.1 NumPy package
2.2 NumPy array
2.3 Linear Algebra

Numerical programming

▶ NumPy package

The *Numerical Python* package `NumPy` provides efficient tools for scientific computing and data analysis:

- `np.array()`: multidimensional array capable of doing fast and efficient computations,
- Built-in mathematical functions on arrays without writing loops,
- Built-in linear algebra functions.

Import NumPy

```
import numpy as np
```

Element-wise addition

```
vec1 = [1, 2, 3, 4, 5, 6, 7, 8, 9]
vec2 = np.array(vec1)
print(vec1 + vec1)

## [1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, 3, 4, 5, 6, 7, 8, 9]

print(vec2 + vec2)

## [ 2  4  6  8 10 12 14 16 18]

for i in range(len(vec1)):
    vec1[i] += vec1[i]
print(vec1)

## [2, 4, 6, 8, 10, 12, 14, 16, 18]
```

Matrix multiplication

```
mat1 = [[1, 2, 3], [4, 5, 6], [7, 8, 9]]
mat2 = np.array(mat1)
print(np.dot(mat2, mat2))

## [[ 30  36  42]
##  [ 66  81  96]
##  [102 126 150]]

mat3 = np.zeros([3, 3])
for i in range(3):
    for k in range(3):
        for j in range(3):
            mat3[i][k] = mat3[i][k] + mat1[i][j] * mat1[j][k]
print(mat3)

## [[ 30.  36.  42.]
##  [ 66.  81.  96.]
##  [102. 126. 150.]]
```

Time comparison

```
import time
mat1 = np.random.rand(50, 50)
mat2 = np.array(mat1)
t = time.time()
mat3 = np.dot(mat2, mat2)
nptime = time.time() - t
mat3 = np.zeros([50, 50])
t = time.time()
for i in range(50):
    for k in range(50):
        for j in range(50):
            mat3[i][k] = mat3[i][k] + mat1[i][j] * mat1[j][k]
pytime = time.time() - t
times = str(pytime / nptime)
print("NumPy is " + times + " times faster!")

## NumPy is 35.166825796371846 times faster!
```

Numerical programming

▶ NumPy array

`np.array(list)`: converts python list into NumPy arrays.
`array.ndim`: returns dimension of the array.
`array.shape`: return shape of the array as a list.

Creation

```
arr1 = [4, 8, 2]
arr1 = np.array(arr1)
arr2 = np.array([24.3, 0., 8.9, 4.4, 1.65, 45])
arr3 = np.array([[4, 8, 5], [9, 3, 4], [1, 0, 6]])
print(arr1.ndim)

## 1

print(arr3.shape)

## (3, 3)
```

From now on, the name array refers to an `np.array()`.

Array creation functions

`np.arange(start, stop, step)`: array of values from `start` to `stop`.
`np.zeros((rows, columns))`: array with all values set to `0`.
`np.identity(dimension)`: identity matrix of a certain `dimension`.

Creation functions

```
print(np.zeros((4, 3)))

## [[0. 0. 0.]
##  [0. 0. 0.]
##  [0. 0. 0.]
##  [0. 0. 0.]]

print(np.arange(6))

## [0 1 2 3 4 5]

print(np.identity(3))

## [[1. 0. 0.]
##  [0. 1. 0.]
##  [0. 0. 1.]]
```

`array.linspace(start, stop, n)`: array of `n` evenly divided values from `start` to `stop`.
`array.full((row, column), k)`: array with all values set to `k`.

Array creation

```
print(np.linspace(0, 80, 5))

## [ 0. 20. 40. 60. 80.]

print(np.full((5, 4), 7))

## [[7 7 7 7]
##  [7 7 7 7]
##  [7 7 7 7]
##  [7 7 7 7]
##  [7 7 7 7]]
```

`np.random.rand(rows, columns)`: array of random floats between zero and one.
`np.rondom.randint(k, size=(rows, columns))`: array of random integers between `0` and `k-1`.

Array of random numbers

```
print(np.random.rand(3, 3))

## [[0.65417053 0.63215654 0.72761157]
##  [0.30757468 0.64874108 0.69997956]
##  [0.74054193 0.57131055 0.77555459]]

print(np.random.randint(10, size=(5, 4)))

## [[4 2 0 4]
##  [3 1 0 9]
##  [9 6 0 0]
##  [3 8 1 9]
##  [9 7 6 7]]
```

Reference

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr = arr3
arr[1, 1] = 777
print(arr3)

## [[  4   8   5]
##  [  9 777   4]
##  [  1   0   6]]

arr3[1, 1] = 3
```

call-by-reference

`arr = arr3` binds `arr` to the existing `arr3`. They both refer to the same object.

`array.copy()`: copy array without reference (call-by-value).

Copy

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr = arr3.copy()
arr[1, 1] = 777
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]
```

Reference

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr = arr3
arr[1, 1] = 777
print(arr3)

## [[  4   8   5]
##  [  9 777   4]
##  [  1   0   6]]

arr3[1, 1] = 3
```

Function	Description
array	Convert input array in NumPy array
arange(start,stop,step)	Creates array from given input
ones	Creates array containing only ones
zeros	Creates array containing only zeros
empty	Allocating memory without specific values
eye, identity	Creates N x N identity matrix
linspace	Creats array of evenly divided values
full	Creates array with values set to one number
random.rand	Creates array of random floats
random.randint	Creates array of random int

`array.dtype`: type of array,
`array.astype(np.type)`: manual typecast.

Data types

```
print(arr1.dtype)

## int64

print(arr2.dtype)

## float64

arr1 = arr1 * 2.5
print(arr1.dtype)

## float64

arr1 = (arr1 / 2.5).astype(np.int64)
print(arr1.dtype)

## int64
```

Element-wise operations

Calculation operators on NumPy arrays operate element-wise.

Element-wise operations

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

print(arr3 + arr3)

## [[ 8 16 10]
##  [18  6  8]
##  [ 2  0 12]]

print(arr3**2)

## [[16 64 25]
##  [81  9 16]
##  [ 1  0 36]]
```

Matrix multiplication

Operator * applied on arrays does not do the matrix multiplication.

Element-wise operations

```
print(arr3 * arr3)

## [[16 64 25]
##  [81  9 16]
##  [ 1  0 36]]

arr = np.ones((3, 2))
print(arr)

## [[1. 1.]
##  [1. 1.]
##  [1. 1.]]

print(arr3 * arr)  # not defined for element-wise multiplication

## ValueError: operands could not be broadcast together
```

`array[start : stop : step]`: Selecting a subset of the data.

Slicing in one dimension

```
arr = np.arange(10)
print(arr)

## [0 1 2 3 4 5 6 7 8 9]

print(arr[4])

## 4

print(arr[3:7])

## [3 4 5 6]

print(arr)

## [0 1 2 3 4 5 6 7 8 9]
```

Slicing in one dimension with steps

```
print(arr[:7])

## [0 1 2 3 4 5 6]

print(arr[-3:])

## [7 8 9]

print(arr[::-1])

## [9 8 7 6 5 4 3 2 1 0]

print(arr[::2])

## [0 2 4 6 8]

print(arr[:5:-1])

## [9 8 7 6]
```

Slicing in higher dimensions

In n-dimensional arrays the element at each index is an (n-1)-dimensional array.

Indexing in two dimensions

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

vec = arr3[1]
print(vec)

## [9 3 4]

print(arr3[1, 0])

## 9
```

Slicing in two dimensions

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

print(arr3[0:2, 0:2])

## [[4 8]
##  [9 3]]

print(arr3[2:, :])

## [[1 0 6]]
```

Slicing

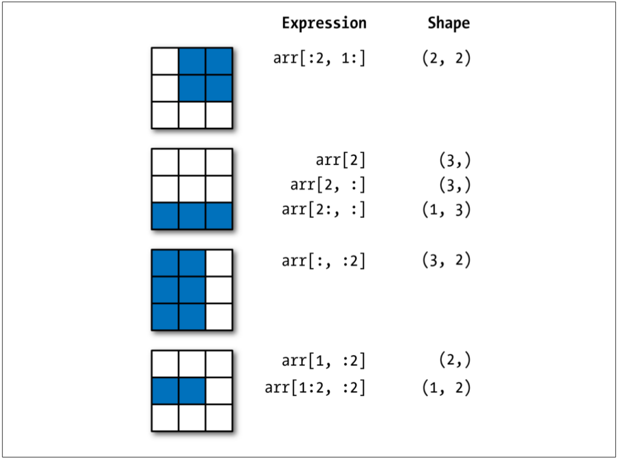


Figure: Python for Data Analysis (2017) on page 99

Views on arrays

So far, selecting by index numbers or slicing belongs to *basic indexing* in NumPy. With basic indexing you get NO COPY of your data but a so-called *view* on the existing data set – a different perspective.
A view on an array can be seen as a reference to a rectangular memory area of its values. The view is intended to

- edit a rectangular part of a matrix, e.g., a sub-matrix, a column, or a single value,
- change the shape of the matrix or the arrangement of its elements, e.g., transpose or reshape a matrix,
- change the visual representation of values, e.g, to cast a `float` array into an `int` array,
- map the values in other program areas.

The crucial point here is that for efficiency reasons data arrays in your working memory do not have to be copied again and again for simple index operations, which would require an excessive additional effort writing to the computer memory.

A view is created automatically when you do basic indexing such as slicing:

Create a view by slicing

```
column = arr3[:, 1]
print(column)

## [8 3 0]

print(column.base)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

column[1] = 100
print(arr3)

## [[  4   8   5]
##  [  9 100   4]
##  [  1   0   6]]
```

Create a view by slicing

```
elem = column[1:2]
print(elem.base)

## [[  4   8   5]
##  [  9 100   4]
##  [  1   0   6]]

elem[0] = 3
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]
```

- The middle column is a view of the base array referenced by `arr3`,
- Any changes to the values of a view directly affect the base data,
- A view of a view is another view on the same base matrix.

In addition, an array contains methods and attributes that return a view of its data:

Obtain a view

```
arr3_t = arr3.T
print(arr3_t)

## [[4 9 1]
##  [8 3 0]
##  [5 4 6]]

print(arr3_t.flags.owndata)

## False

arr3_r = arr3.reshape(1, 9)
print(arr3_r)

## [[4 8 5 9 3 4 1 0 6]]

print(arr3_t.flags.owndata)

## False
```

Obtain a view

```
arr3_v = arr3.view()
print(arr3_v.flags.owndata)

## False
```

- The transposed matrix is a predefined view that is available as an attribute,
- Reshaping is also just another way of looking at the same set of data,
- By means of the method `view()` you create a view with an identical representation.

The behavior described above changes with *advanced indexing*, i. e., if at least one component of the index tuple is not a scalar index number or slice. The case of *fancy indexing* is described below:

Advanced and basic indexing

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr = arr3[[0, 2], [0, 2]]
print(arr)

## [4 6]

print(arr.base)

## None
```

Advanced and basic indexing

```
arr = arr3[0:3:2, 0:3:2]
print(arr)

## [[4 5]
##  [1 6]]

print(arr.base)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]
```

- Contrary to intuition, fancy indexing does not return a (2×2)-matrix, but a vector of the matrix elements $(0, 0)$ and $(2, 2)$. This is a complete copy – a new object and not a view to the original matrix.
- A submatrix (view) with the corner elements of the initial matrix can be obtained with slicing.

Filter arrays without using loops by conditional indexing.

Find and replace values in arrays, condition: smaller

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr = arr3.copy()
arr[arr < 5] = 0
print(arr)

## [[0 8 5]
##  [9 0 0]
##  [0 0 6]]
```

Find and replace values in arrays, condition: equal

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr = arr3.copy()
arr[arr == 4] = 100
print(arr)

## [[100   8   5]
##  [  9   3 100]
##  [  1   0   6]]
```

`array.reshape((rows, columns))`: reshaping existing array.
`array.resize((rows, columns))`: changes array shape to `rows` x `columns` and fills new values with `0`.

Reshape

```
arr = np.arange(15)
print(arr.reshape((3, 5)))

## [[ 0  1  2  3  4]
##  [ 5  6  7  8  9]
##  [10 11 12 13 14]]

arr.resize((3, 7))
print(arr)

## [[ 0  1  2  3  4  5  6]
##  [ 7  8  9 10 11 12 13]
##  [14  0  0  0  0  0  0]]
```

`np.append(array, value)`: appends value to the end of array.
`np.insert(array, index, value)`: inserts values before index.
`np.delete(array, index, axis)`: deletes row or column on index.

Naming

```
a = np.arange(5)
a = np.append(a, 8)
a = np.insert(a, 3, 77)
print(a)

## [ 0  1  2 77  3  4  8]

a.resize((3, 3))
print(np.delete(a, 1, axis=0))

## [[0 1 2]
##  [8 0 0]]
```

`np.concatenate((arr1, arr2), axis)`: join a sequence of arrays along an existing axis.
`np.split(array, n)`: split an array into multiple sub-arrays.
`np.hsplit(array, n)`: split an array into multiple sub-arrays horizontally.

Naming

```
print(np.concatenate((a, np.arange(6).reshape(2, 3)), axis=0))

## [[ 0  1  2]
##  [77  3  4]
##  [ 8  0  0]
##  [ 0  1  2]
##  [ 3  4  5]]

print(np.split(np.arange(8), 4))

## [array([0, 1]), array([2, 3]), array([4, 5]), array([6, 7])]
```

`array.T`: transposed array (as a view).

Transpose

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

print(arr3.T)

## [[4 9 1]
##  [8 3 0]
##  [5 4 6]]

print(np.eye(3).T)

## [[1. 0. 0.]
##  [0. 1. 0.]
##  [0. 0. 1.]]
```

`np.dot(array1, array2)`: matrix multiplication of `array1` and `array2`.

Matrix multiplication

```
res = np.dot(arr3, np.arange(18).reshape((3, 6)))
print(res)

## [[108 125 142 159 176 193]
##  [ 66  82  98 114 130 146]
##  [ 72  79  86  93 100 107]]

res = np.dot(np.eye(4), np.arange(16).reshape((4, 4)))
print(res)

## [[ 0.  1.  2.  3.]
##  [ 4.  5.  6.  7.]
##  [ 8.  9. 10. 11.]
##  [12. 13. 14. 15.]]
```

Element-wise functions

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

print(np.sqrt(arr3))

## [[2.         2.82842712 2.23606798]
##  [3.         1.73205081 2.        ]
##  [1.         0.         2.44948974]]

print(np.exp(arr3))

## [[5.45981500e+01 2.98095799e+03 1.48413159e+02]
##  [8.10308393e+03 2.00855369e+01 5.45981500e+01]
##  [2.71828183e+00 1.00000000e+00 4.03428793e+02]]
```

Overview: element-wise array functions

Function	Description
abs	Absolute value of integer and floating point
sqrt	Sqare root
exp	Exponential function
log, log10, log2	Natural logarithm, log base 10, log base 2
sign	Sign (1 : positiv, 0: zero, -1 : negative)
ceil	Rounding up to integer
floor	Round down to integer
rint	Round to nearest integer
modf	Returns fractional parts
sin, cos, tan, sinh, cosh, tanh, arcsin, ...	

Binary

```
x = np.array([3, -6, 8, 4, 3, 5])
y = np.array([3, 5, 7, 3, 5, 9])
print(np.maximum(x, y))

## [3 5 8 4 5 9]

print(np.greater_equal(x, y))

## [ True False  True  True False False]

print(np.add(x, y))

## [ 6 -1 15  7  8 14]

print(np.mod(x, y))

## [0 4 1 1 3 5]
```

Function	Description
add	Add elements of arrays
subtract	Subtract elements in the second from the first array
multiply	Multiply elements
divide	Divide elements
power	Raise elements in first array to powers in second
maximum	Element-wise maximum
minimum	Element-wise minimum
mod	Element-wise modulus
greater, less, equal gives boolean	

`np.meshgrid(array1, array2)`: coordinate matrice from coordinate arrays.

Evaluate the function $f(x, y) = \sqrt{x^2 + y^2}$ on a 10 x 10 grid

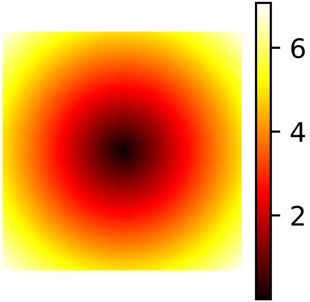
```
p = np.arange(-5, 5, 0.01)
x, y = np.meshgrid(p, p)
print(x)

## [[-5.   -4.99 -4.98 ...  4.97  4.98  4.99]
##  [-5.   -4.99 -4.98 ...  4.97  4.98  4.99]
##  [-5.   -4.99 -4.98 ...  4.97  4.98  4.99]
##  ...
##  [-5.   -4.99 -4.98 ...  4.97  4.98  4.99]
##  [-5.   -4.99 -4.98 ...  4.97  4.98  4.99]
##  [-5.   -4.99 -4.98 ...  4.97  4.98  4.99]]
```

Evaluate the function $f(x, y) = \sqrt{x^2 + y^2}$ on a 10 x 10 grid.

```
import matplotlib.pyplot as plt
val = np.sqrt(x**2 + y**2)
plt.figure(figsize=(2, 2))
plt.imshow(val, cmap="hot")
plt.colorbar()
```

Evaluate the function $f(x, y) = \sqrt{x^2 + y^2}$ on a 10 x 10 grid.

```
plt.show()
```



`np.where(condition, a, b)`: If condition is `True`, take value from `a`, else take `b`.

Conditional logic

```
a = np.array([4, 7, 5, -7, 9, 0])
b = np.array([-1, 9, 8, 3, 3, 3])
cond = np.array([True, True, False, True, False, False])
res = np.where(cond, a, b)
print(res)

## [ 4  7  8 -7  3  3]

res = np.where(a <= b, b, a)
print(res)

## [4 9 8 3 9 3]
```

Conditional logic, examples

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

res = np.where(arr3 < 5, 0, arr3)
print(res)

## [[0 8 5]
##  [9 0 0]
##  [0 0 6]]

even = np.where(arr3 % 2 == 0, arr3, arr3 + 1)
print(even)

## [[ 4  8  6]
##  [10  4  4]
##  [ 2  0  6]]
```

`array.mean()`: mean of all array elements.
`array.sum()`: sum of all array elements.

Statistical methods

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

print(arr3.mean())

## 4.444444444444445

print(arr3.sum())

## 40

print(arr3.argmin())

## 7
```

Method	**Description**
sum	Sum of all array elements
mean	Mean of all array elements
std, var	Standard deviation, variance
min, max	Minimum and Maximum value in array
argmin, argmax	Indices of Minimum and Maximum value

Axes are defined for arrays with more than one dimension. A two-dimensional array has two axes. The first one is running vertically downwards across the rows (`axis=0`), the second one running horizontally across the columns (`axis=1`).

Axis

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

print(arr3.sum(axis=0))

## [14 11 15]

print(arr3.sum(axis=1))

## [17 16  7]
```

`array.sort(axis)`: sort array by an `axis`.

Sorting one-dimensional arrays

```
print(arr2)

## [24.3   0.    8.9   4.4   1.65 45.  ]

arr2.sort()
print(arr2)

## [ 0.    1.65  4.4   8.9  24.3  45.  ]
```

Sorting two-dimensional arrays

```
print(arr3)

## [[4 8 5]
##  [9 3 4]
##  [1 0 6]]

arr3.sort()
print(arr3)

## [[4 5 8]
##  [3 4 9]
##  [0 1 6]]

arr3.sort(axis=0)
print(arr3)

## [[0 1 6]
##  [3 4 8]
##  [4 5 9]]
```

The default axis using `sort()` is −1, which means to sort along the last axis (in this case axis 1).

Numerical programming

▶ Linear Algebra

Import numpy.linalg

```
import numpy.linalg as nplin
```

`nplin.inv(array)`: inverse matrix.
`np.allclose(array1, array2)`: returns `True` if two arrays are element-wise equal within a tolerance.

Inverse

```
inv = nplin.inv(arr3)
print(inv)

## [[  4. -21.  16.]
##  [ -5.  24. -18.]
##  [  1.  -4.   3.]]

print(np.allclose(np.identity(3), np.dot(inv, arr3)))

## True
```

`nplin.det(array)`: compute determininat.
`np.trace(array)`: compute trace.
`np.diag(array)`: return diagonal elements as an array.

Linear algebra functions

```
print(nplin.det(arr3))

## -1.0

print(np.trace(arr3))

## 13

print(np.diag(arr3))

## [0 4 9]
```

`nplin.eig(array)`: return array of eigenvalues and array of eigenvectors as a list.

Get eigenvalues and eigenvectors

```
A = np.array([[3, -1, 0], [2, 0, 0], [-2, 2, -1]])
eigenval, eigenvec = nplin.eig(A)
print(eigenval)

## [-1.  1.  2.]

print(eigenvec)

## [[ 0.         -0.40824829 -0.70710678]
##  [ 0.         -0.81649658 -0.70710678]
##  [ 1.         -0.40824829  0.        ]]
```

Check eigenvalues and eigenvectors

```
print(eigenval * eigenvec)

## [[-0.          -0.40824829 -1.41421356]
##  [-0.          -0.81649658 -1.41421356]
##  [-1.          -0.40824829  0.        ]]

print(np.dot(A, eigenvec))

## [[ 0.          -0.40824829 -1.41421356]
##  [ 0.          -0.81649658 -1.41421356]
##  [-1.          -0.40824829  0.        ]]
```

$$\begin{pmatrix} 3 & -1 & 0 \\ 2 & 0 & 0 \\ -2 & 2 & -1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = (-1) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}$$

`nplin.qr(array)`: QR decomposition, returns Q and R as lists.

QR decomposition

```
Q, R = nplin.qr(arr3)
print(Q)

## [[ 0.          0.98058068  0.19611614]
##  [-0.6         0.15689291 -0.78446454]
##  [-0.8        -0.11766968  0.58834841]]

print(R)

## [[ -5.          -6.4        -12.        ]
##  [  0.           1.0198039    6.07960019]
##  [  0.           0.           0.19611614]]

print(np.allclose(arr3, np.dot(Q, R)))

## True
```

`nplin.solve(A, b)`: return solution of the linearsystem $Ax = b$.

Solve linearsystems

```
b = np.array([7, 4, 8])
x = nplin.solve(A, b)
print(x)

## [  2.  -1. -14.]

print(np.allclose(np.dot(A, x), b))

## True
```

$$\begin{aligned} 3x_1 - 1x_2 + 0x_3 &= 7 \\ 2x_1 - 0x_2 + 0x_3 &= 4 \\ -2x_1 + 2x_2 - 1x_3 &= 8 \end{aligned} \rightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ -14 \end{pmatrix}$$

Function	Description
np.dot	Matrix multiplication
np.trace	Sum of the diagonal elements
np.diag	Diagonal elements as an array
nplin.det	Matrix determinant
nplin.eig	Eigenvalues and eigenvectors
nplin.inv	Inverse matrix
nplin.qr	QR decomposition
nplin.solve	Solve linearsystem

Data formats and handling

Data formats and handling

▶ Pandas

Pandas

The package `pandas` is a free software library for Python including the following features:

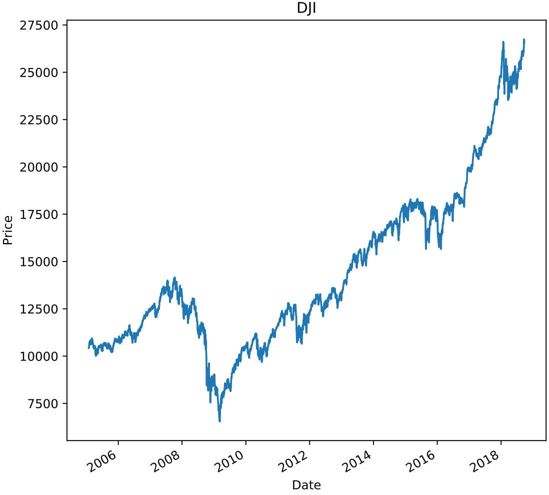
- Data manipulation and analysis,
- DataFrame objects and Series,
- Export and import data from files and web,
- Handling of missing data.

→ Provides high-performance data structures and data analysis tools.

With pandas you can import and visualize financial data in only a few lines of code.

Motivation

```python
import pandas as pd
import matplotlib.pyplot as plt
fig = plt.figure()
ax = fig.add_subplot(1, 1, 1)
dow = pd.read_csv("data/dji.csv", index_col=0, parse_dates=True)
close = dow["Close"]
close.plot(ax=ax)
ax.set_xlabel("Date")
ax.set_ylabel("Price")
ax.set_title("DJI")
fig.savefig("out/dji.pdf", format="pdf")
```



DJI
Price
27500
25000
22500
20000
17500
15000
12500
10000
7500
2006
2008
2010
2012
2014
2016
2018
Date

Data formats and handling

▶ Series

Series are a data structure in pandas.

- One-dimensional array-like object,
- Containing a sequence of values and an corresponding array of labels, called the index,
- The string representation of a Series displays the index an the right and the values on the right,
- The default index consists of the integers 0 through N-1.

String representation of a Series

```
## 0     3
## 1     7
## 2    -8
## 3     4
## 4    26
## dtype: int64
```

Import pandas

```
import pandas as pd
```

`pd.Series()`: one-dimensional array-like object including values and an index.

Series

```
obj = pd.Series([2, -5, 9, 4])
print(obj)

## 0    2
## 1   -5
## 2    9
## 3    4
## dtype: int64
```

- Simple Series formed only from a list,
- Index is added automatically.

Series indexing vs. Numpy indexing

```
obj2 = pd.Series([2, -5, 9, 4], index=["a", "b", "c", "d"])
npobj = np.array([2, -5, 9, 4])
print(obj2)

## a    2
## b   -5
## c    9
## d    4
## dtype: int64

print(obj2["b"])

## -5

print(npobj[1])

## -5
```

- NumPy arrays can only be indexed by integers while Series can be indexed by the manually set index.

Series creation from Numpy arrays

```
npobj = np.array([2, -5, 9, 4])
obj2 = pd.Series(npobj, index=["a", "b", "c", "d"])
print(obj2)

## a    2
## b   -5
## c    9
## d    4
## dtype: int64
```

Pandas Series can be created from:

- Lists,
- NumPy arrays,
- Dicts.

- The index of the Series can be set manually,
- Compared to NumPy array you can use the set index to select single values,
- Data contained in a dict can be passed to a Series. The index of the resulting Series consists of the dict's keys.

Series from dicts

```
dictdata = {"Göttingen": 117665, "Northeim": 28920,
            "Hannover": 532163, "Berlin": 3574830}
obj3 = pd.Series(dictdata)
print(obj3)

## Göttingen     117665
## Northeim       28920
## Hannover      532163
## Berlin       3574830
## dtype: int64
```

Dict to Series with manual index

```
cities = ["Hamburg", "Göttingen", "Berlin", "Hannover"]
obj4 = pd.Series(dictdata, index=cities)
print(obj4)

## Hamburg              NaN
## Göttingen       117665.0
## Berlin         3574830.0
## Hannover        532163.0
## dtype: float64
```

- Passing a dict to a Series, the index can be set manually,
- `NaN` (not a number) marks missing values where the index and the dict do not match.

`Series.values`: returns the values of a Series.
`Series.index`: returns the index of a Series.

Series properties

```
print(obj.values)

## [ 2 -5  9  4]

print(obj.index)

## RangeIndex(start=0, stop=4, step=1)

print(obj2.index)

## Index(['a', 'b', 'c', 'd'], dtype='object')
```

- The values and the index of a DataFrame can be printed separately as a list,
- The default index is given as a `RangeIndex`.

Series manipulation

```
print(obj2[["c", "d", "a"]])

## c    9
## d    4
## a    2
## dtype: int64

print(obj2[obj2 < 0])

## b   -5
## dtype: int64
```

NumPy-like functions can be applied on Series

- For filtering data,
- To do scalar multiplications or applying math functions,
- The index-value link will be preserved.

Series functions

```
print(obj2 * 2)

## a     4
## b   -10
## c    18
## d     8
## dtype: int64

print(np.exp(obj2)["a":"c"])

## a       7.389056
## b       0.006738
## c    8103.083928
## dtype: float64

print("c" in obj2)

## True
```

- Mathematical functions on a Series will only be applied on the values not on its index.

Series manipulation

```
obj4["Hamburg"] = 1900000
print(obj4)

## Hamburg      1900000.0
## Göttingen     117665.0
## Berlin       3574830.0
## Hannover      532163.0
## dtype: float64

obj4[["Berlin", "Hannover"]] = [3600000, 1100000]
print(obj4)

## Hamburg      1900000.0
## Göttingen     117665.0
## Berlin       3600000.0
## Hannover     1100000.0
## dtype: float64
```

- Values can be manipulated by using the labels in the index,
- Sets of values can be set in one line.

`pd.isnull()`: `True` if data is missing.
`pd.notnull()`: `False` if data is missing.

NaN

```
print(pd.isnull(obj4))

## Hamburg      False
## Göttingen    False
## Berlin       False
## Hannover     False
## dtype: bool

print(pd.notnull(obj4))

## Hamburg      True
## Göttingen    True
## Berlin       True
## Hannover     True
## dtype: bool
```

There are not two values to align for `Hamburg` and `Northeim` so they are marked with `NaN` (not a number).

Data 1

```
print(obj3)

## Göttingen      117665
## Northeim        28920
## Hannover       532163
## Berlin        3574830
## dtype: int64
```

Data 2

```
print(obj4)

## Hamburg       1900000.0
## Göttingen      117665.0
## Berlin        3600000.0
## Hannover      1100000.0
## dtype: float64
```

Align data

```
print(obj3 + obj4)

## Berlin       7174830.0
## Göttingen     235330.0
## Hamburg            NaN
## Hannover     1632163.0
## Northeim           NaN
## dtype: float64
```

`Series.name`: name of the Series
`Series.index.name`: name of the index

Naming

```
obj4.name = "population"
obj4.index.name = "city"
print(obj4)

## city
## Hamburg      1900000.0
## Göttingen     117665.0
## Berlin       3600000.0
## Hannover     1100000.0
## Name: population, dtype: float64
```

- The attribute `name` will change the name of the existing Series,
- There is no default name of the Series or the index.

- NumPy arrays are accessed by their integer position.
- Series can definied and accessed by your own index, including letters and numbers.
- Different Series can be aligned efficiently by the index.
- Series can work with missing values, so operations do not automatically fail.

Data formats and handling

▶ DataFrame

DataFrame

- *DataFrames* are the primary structure of pandas,
- It represents a table of data with an ordered collection of columns,
- Each column can have a different data type,
- A DataFrame can be thought of as a dict of Series sharing the same index,
- Physically a DataFrame is two-dimensional, but by using hierachical indexing it can respresent higher dimensional data.

String representation of a DataFrame

```
##     company   price   volume
## 0   Daimler   69.20  4456290
## 1      E.ON    8.11  3667975
## 2   Siemens  110.92  3669487
## 3      BASF   87.28  1778058
## 4       BMW   87.81  1824582
```

pd.DataFrame(): a DataFrame is a tabular-like structure. It is two-dimensional and has labeled axis (rows and columns).

Creating a DataFrame

```
data = {"company": ["Daimler", "E.ON", "Siemens", "BASF", "BMW"],
        "price": [69.2, 8.11, 110.92, 87.28, 87.81],
        "volume": [4456290, 3667975, 3669487, 1778058, 1824582]}
frame = pd.DataFrame(data)
print(frame)

##    company   price   volume
## 0  Daimler   69.20  4456290
## 1     E.ON    8.11  3667975
## 2  Siemens  110.92  3669487
## 3     BASF   87.28  1778058
## 4      BMW   87.81  1824582
```

- In this example the construction of the DataFrame `frame` is done by passing a dict of equal-length lists,
- Instead of passing a dict of lists, it is also possible to pass a dict of NumPy arrays.

Print DataFrame

```
frame2 = pd.DataFrame(data, columns=["company", "volume",
                                     "price", "change"])
print(frame2)

##    company   volume   price change
## 0  Daimler  4456290   69.20    NaN
## 1     E.ON  3667975    8.11    NaN
## 2  Siemens  3669487  110.92    NaN
## 3     BASF  1778058   87.28    NaN
## 4      BMW  1824582   87.81    NaN
```

- Passing a column that is not contained in the dict, it will be marked with `NaN`,
- The default index will be assigned automatically as with Series.

Type	Description
2D NumPy arrays	A matrix of data
dict of arrays, lists, or tuples	Each sequence becomes a column
dict of Series	Each value becomes a column
dict of dicts	Each inner dict becomes a column
List of dicts or Series	Each item becomes a row
List of lists or tuples	Treated as the 2D NumPy arrays
Another DataFrame	Same indexes

Add data to DataFrame

```
frame2["change"] = [1.2, -3.2, 0.4, -0.12, 2.4]
print(frame2["change"])

## 0    1.20
## 1   -3.20
## 2    0.40
## 3   -0.12
## 4    2.40
## Name: change, dtype: float64
```

- Selecting the column of DataFrame, a Series is returned,
- A attribute-like access, e. g., `frame2.change`, is also possible,
- The returned Series has the same index as the initial DataFrame.

Indexing DataFrames

```
print(frame2[["company", "change"]])

##    company  change
## 0  Daimler    1.20
## 1     E.ON   -3.20
## 2  Siemens    0.40
## 3     BASF   -0.12
## 4      BMW    2.40
```

- Using a list of multiple columns while indexing, the result is a DataFrame,
- The returned DataFrame has the same index as the initial one.

`del DataFrame[column]`: delete column from DataFrame.

DataFrame delete column

```
del frame2["volume"]
print(frame2)

##    company   price  change
## 0  Daimler   69.20    1.20
## 1     E.ON    8.11   -3.20
## 2  Siemens  110.92    0.40
## 3     BASF   87.28   -0.12
## 4      BMW   87.81    2.40

print(frame2.columns)

## Index(['company', 'price', 'change'], dtype='object')
```

Naming properties

```
frame2.index.name = "number:"
frame2.columns.name = "feature:"
print(frame2)

## feature:  company   price  change
## number:
## 0         Daimler   69.20    1.20
## 1            E.ON    8.11   -3.20
## 2         Siemens  110.92    0.40
## 3            BASF   87.28   -0.12
## 4             BMW   87.81    2.40
```

- In DataFrames there is no default name for the index or the columns.

`DataFrame.reindex()`: creates new DataFrame with data conformed to a new index, the initial DataFrame will not be changed.

Reindexing

```
frame3 = frame.reindex([0, 2, 3, 4])
print(frame3)

##    company   price   volume
## 0  Daimler   69.20  4456290
## 2  Siemens  110.92  3669487
## 3     BASF   87.28  1778058
## 4      BMW   87.81  1824582
```

- Index values that are not already present will be filled with `NaN` by default,
- There are many options for filling missing values.

Filling missing values

```
frame4 = frame.reindex(index=[0, 2, 3, 4, 5], fill_value=0,
                       columns=["company", "price", "market cap"])
print(frame4)

##     company   price  market cap
## 0   Daimler   69.20           0
## 2   Siemens  110.92           0
## 3      BASF   87.28           0
## 4       BMW   87.81           0
## 5         0    0.00           0

frame4 = frame.reindex(index=[0, 2, 3, 4], fill_value=np.nan,
                       columns=["company", "price", "market cap"])
print(frame4)

##     company   price  market cap
## 0   Daimler   69.20         NaN
## 2   Siemens  110.92         NaN
## 3      BASF   87.28         NaN
## 4       BMW   87.81         NaN
```

`DataFrame.fillna(value)`: filling `NaN` with value

Filling NaN

```
print(frame4[:3])

##    company   price  market cap
## 0  Daimler   69.20         NaN
## 2  Siemens  110.92         NaN
## 3     BASF   87.28         NaN

frame4.fillna(1000000, inplace=True)
print(frame4[:3])

##    company   price  market cap
## 0  Daimler   69.20   1000000.0
## 2  Siemens  110.92   1000000.0
## 3     BASF   87.28   1000000.0
```

- The option `inplace=True` fills the current DafaFrame (here `frame4`). Without using `inplace` a new DataFrame will be created, filled with `NaN` values.

`DataFrame.drop(index, axis)`: returns a new object with labels in requested axis removed.

Dropping index

```
frame5 = frame
print(frame5)

##    company   price   volume
## 0  Daimler   69.20  4456290
## 1     E.ON    8.11  3667975
## 2  Siemens  110.92  3669487
## 3     BASF   87.28  1778058
## 4      BMW   87.81  1824582

print(frame5.drop([1, 2]))

##    company  price   volume
## 0  Daimler  69.20  4456290
## 3     BASF  87.28  1778058
## 4      BMW  87.81  1824582
```

Dropping column

```
print(frame5[:2])

##    company  price   volume
## 0  Daimler  69.20  4456290
## 1     E.ON   8.11  3667975

print(frame5.drop("price", axis=1)[:3])

##    company   volume
## 0  Daimler  4456290
## 1     E.ON  3667975
## 2  Siemens  3669487

print(frame5.drop(2, axis=0))

##    company  price   volume
## 0  Daimler  69.20  4456290
## 1     E.ON   8.11  3667975
## 3     BASF  87.28  1778058
## 4      BMW  87.81  1824582
```

- Indexing of DataFrames works like indexing an `numpy` array, you can use the default index values and a manually set index.

Indexing

```
print(frame)

##     company   price   volume
## 0   Daimler   69.20  4456290
## 1      E.ON    8.11  3667975
## 2   Siemens  110.92  3669487
## 3      BASF   87.28  1778058
## 4       BMW   87.81  1824582

print(frame[2:])

##     company   price   volume
## 2   Siemens  110.92  3669487
## 3      BASF   87.28  1778058
## 4       BMW   87.81  1824582
```

Indexing

```
frame6 = pd.DataFrame(data, index=["a", "b", "c", "d", "e"])
print(frame6)

##    company   price   volume
## a  Daimler   69.20  4456290
## b     E.ON    8.11  3667975
## c  Siemens  110.92  3669487
## d     BASF   87.28  1778058
## e      BMW   87.81  1824582

print(frame6["b":"d"])

##    company   price   volume
## b     E.ON    8.11  3667975
## c  Siemens  110.92  3669487
## d     BASF   87.28  1778058
```

- When *slicing with labels* the end element is inclusive.

`DataFrame.loc()`: select a subset of rows and columns from a DataFrame using axis labels.
`DataFrame.iloc()`: select a subset of rows and columns from a DataFrame using integers.

Selection with loc and iloc

```
print(frame6.loc["c", ["company", "price"]])

## company    Siemens
## price       110.92
## Name: c, dtype: object

print(frame6.iloc[2, [0, 1]])

## company    Siemens
## price       110.92
## Name: c, dtype: object
```

Selection with loc and iloc

```
print(frame6.loc[["c", "d", "e"], ["volume", "price", "company"]])

##     volume   price  company
## c  3669487  110.92  Siemens
## d  1778058   87.28     BASF
## e  1824582   87.81      BMW

print(frame6.iloc[2:, ::-1])

##     volume   price  company
## c  3669487  110.92  Siemens
## d  1778058   87.28     BASF
## e  1824582   87.81      BMW
```

- Both of the indexing functions work with slices or lists of labels,
- Many ways to select and rearrange pandas objects.

Type	Description
df[val]	Select single column or set of columns
df.loc[val]	Select single row or set of rows
df.loc[:, val]	Select single column or set of columns
df.loc[val1, val2]	Select row and column by label
df.iloc[where]	Select row or set of rows by integer position
df.iloc[:, where]	Select column or set of columns by integer pos.
df.iloc[w1, w2]	Select row and column by integer position

- Hierarchical indexing enables you to have multiple index levels.

Multiindex

```
ind = [["a", "a", "a", "b", "b"], [1, 2, 3, 1, 2]]
frame6 = pd.DataFrame(np.arange(15).reshape((5, 3)),
                      index=ind,
                      columns=["first", "second", "third"])
print(frame6)

##       first  second  third
## a 1       0       1      2
##   2       3       4      5
##   3       6       7      8
## b 1       9      10     11
##   2      12      13     14

frame6.index.names = ["index1", "index2"]
print(frame6.index)

## MultiIndex(levels=[['a', 'b'], [1, 2, 3]],
##            labels=[[0, 0, 0, 1, 1], [0, 1, 2, 0, 1]],
##            names=['index1', 'index2'])
```

Selecting of a multiindex

```
print(frame6.loc["a"])

##          first  second  third
## index2
## 1            0       1      2
## 2            3       4      5
## 3            6       7      8

print(frame6.loc["b", 1])

## first      9
## second    10
## third     11
## Name: (b, 1), dtype: int64
```

Series and DataFrames

```
frame7 = frame[["price", "volume"]]
frame7.index = ["Daimler", "E.ON", "Siemens", "BASF", "BMW"]
series = frame7.iloc[2]
print(frame7)

##           price   volume
## Daimler   69.20  4456290
## E.ON       8.11  3667975
## Siemens  110.92  3669487
## BASF      87.28  1778058
## BMW       87.81  1824582

print(series)

## price          110.92
## volume     3669487.00
## Name: Siemens, dtype: float64
```

- Here the Series was generated from the first row of the DataFrame.

Operations between Series and DataFrames down the rows

```
print(frame7 + series)

##            price     volume
## Daimler  180.12  8125777.0
## E.ON     119.03  7337462.0
## Siemens  221.84  7338974.0
## BASF     198.20  5447545.0
## BMW      198.73  5494069.0
```

- By default arithmetic operations between DataFrames and Series match the index of the Series on the DataFrame's columns,
- The operations will be broadcasted along the rows.

Operations between Series and DataFrames down the columns

```
series2 = frame7["price"]
print(frame7.add(series2, axis=0))

##           price      volume
## Daimler  138.40  4456359.20
## E.ON      16.22  3667983.11
## Siemens  221.84  3669597.92
## BASF     174.56  1778145.28
## BMW      175.62  1824669.81
```

- Here, the Series was generated from the `price` column,
- The arithmetic operation will be broadcasted along a column matching the DataFrame's row index (`axis=0`).

Pandas vs Numpy

```
nparr = np.arange(12.).reshape((3, 4))
row = nparr[0]
print(nparr-row)

## [[0. 0. 0. 0.]
##  [4. 4. 4. 4.]
##  [8. 8. 8. 8.]]
```

- Operations between DataFrames are similar to operations between one- and two-dimensional Numpy arrays,
- As in DataFrames and Series the arithmetic operations will be broadcasted along the rows.

`DataFrame.apply(np.function, axis)`: applies a NumPy function on the DataFrame axis.
See also statistical and mathematical NumPy functions.

Numpy functions on DataFrames

```
print(frame7[:2])

##          price   volume
## Daimler  69.20  4456290
## E.ON      8.11  3667975

print(frame7.apply(np.mean))

## price           72.664
## volume     3079278.400
## dtype: float64

print(frame7.apply(np.sqrt)[:2])

##              price       volume
## Daimler  8.318654  2110.992657
## E.ON     2.847806  1915.195812
```

`DataFrame.groupby(col1, col2)`: group DataFrame by columns (grouping by one or more than two columns is also possible).
See also how to import data from CSV files.

Groupby

```
vote = pd.read_csv("data/vote.csv")[["Party", "Member", "Vote"]]
print(vote.head())

##       Party     Member    Vote
## 0   CDU/CSU   Abercron     yes
## 1   CDU/CSU     Albani     yes
## 2   CDU/CSU  Altenkamp     yes
## 3   CDU/CSU  Altmaier   absent
## 4   CDU/CSU     Amthor     yes
```

Adding the functions `count()` or `mean()` to `groupby()` returns the sum or the mean of the grouped columns.

Groupby

```
res = vote.groupby(["Party", "Vote"]).count()
print(res)
```

```
##                          Member
## Party        Vote
## AfD          absent           6
##              no              86
## BÜ90/GR      absent           9
##              no              58
## CDU/CSU      absent           7
##              yes            239
## DIE LINKE.   absent           7
##              no              62
## FDP          absent           5
##              no              75
## Fraktionslos absent           1
##              no               1
## SPD          absent           6
##              yes            147
```

Data formats and handling

- Import/Export data

ex1.csv

```
a, b, c, d, hello
1, 2, 3, 4, world
5, 6, 7, 8, python
2, 3, 5, 7, pandas
```

`pd.read_csv("file")`: read CSV into DataFrame.

Read comma-separated values

```
df = pd.read_csv("data/ex1.csv")
print(df)

##    a  b  c  d   hello
## 0  1  2  3  4   world
## 1  5  6  7  8  python
## 2  2  3  5  7  pandas
```

tab.txt

```
a| b| c| d| hello
1| 2| 3| 4| world
5| 6| 7| 8| python
2| 3| 5| 7| pandas
```

`pd.read_table("file", sep)`: read table with any seperators into DataFrame.

Read table values

```
df = pd.read_table("data/tab.txt", sep="|")
print(df)

##    a   b   c   d    hello
## 0  1   2   3   4    world
## 1  5   6   7   8   python
## 2  2   3   5   7   pandas
```

ex2.csv

```
1, 2, 3, 4, world
5, 6, 7, 8, python
2, 3, 5, 7, pandas
```

CSV file without header row:

Read CSV and header settings

```
df = pd.read_csv("data/ex2.csv", header=None)
print(df)

##    0  1  2  3       4
## 0  1  2  3  4   world
## 1  5  6  7  8  python
## 2  2  3  5  7  pandas
```

ex2.csv

```
1, 2, 3, 4, world
5, 6, 7, 8, python
2, 3, 5, 7, pandas
```

Specify header:

Read CSV and header names

```
df = pd.read_csv("data/ex2.csv",
                 names=["a", "b", "c", "d", "hello"])
print(df)

##    a  b  c  d   hello
## 0  1  2  3  4   world
## 1  5  6  7  8  python
## 2  2  3  5  7  pandas
```

ex2.csv

```
1, 2, 3, 4, world
5, 6, 7, 8, python
2, 3, 5, 7, pandas
```

Use `hello`-column as the index:

Read CSV and specify index

```
df = pd.read_csv("data/ex2.csv",
                 names=["a", "b", "c", "d", "hello"],
                 index_col="hello")
print(df)

##          a  b  c  d
## hello
##  world   1  2  3  4
##  python  5  6  7  8
##  pandas  2  3  5  7
```

ex3.csv

```
1, 2, 3, 4, world
#+#-.,.-'*'-.,
5, 6, 7, 8, python
87646756754456978
2, 3, 5, 7, pandas
```

Skip rows while reading:

Read CSV and choose rows

```
df = pd.read_csv("data/ex3.csv", skiprows=[1, 3])
print(df)

##    1   2   3   4    world
## 0  5   6   7   8   python
## 1  2   3   5   7   pandas
```

`DataFrame.to_csv("filename')`: writing DataFrame to CSV.

Write to CSV

```
df = pd.read_csv("data/ex3.csv", skiprows=[1, 3])
df.to_csv("out/out1.csv")
```

out1.csv

```
,1, 2, 3, 4, world
0,5,6,7,8, python
1,2,3,5,7, pandas
```

In the .csv file, the index and header is included (reason why `,1`).

Write to CSV and settings

```
df = pd.read_csv("data/ex3.csv", skiprows=[1, 3])
df.to_csv("out/out2.csv", index=False, header=False)
```

out2.csv

```
5,6,7,8, python
2,3,5,7, pandas
```

Write to CSV and specify header

```
df = pd.read_csv("data/ex3.csv", skiprows=[1, 3, 4])
df.to_csv("out/out3.csv", index=False,
          header=["a", "b", "c", "d", "e"])
```

out3.csv

```
a,b,c,d,e
5,6,7,8, python
```

Reading Excel files

`pd.read_excel("file.xls")`: read .xls files.

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Adj Close	Volume
2	2018-01-31	1170.569946	1173	1159.130005	1169.939941	1169.939941	1538700
3	2018-02-01	1162.609985	1174	1157.52002	1167.699951	1167.699951	2412100
4	2018-02-02	1122	1123.069946	1107.277954	1111.900024	1111.900024	4857900
5	2018-02-05	1090.599976	1110	1052.030029	1055.800049	1055.800049	3798300
6	2018-02-06	1027.180054	1081.709961	1023.137024	1080.599976	1080.599976	3448000
7	2018-02-07	1081.540039	1081.780029	1048.26001	1048.579956	1048.579956	2341700

Figure: goog.xls

Reading Excel

```
xls_frame = pd.read_excel("data/goog.xls")
```

Excel as a DataFrame

```
print(xls_frame[["Adj Close", "Volume", "High"]])

##       Adj Close   Volume         High
## 0   1169.939941  1538700  1173.000000
## 1   1167.699951  2412100  1174.000000
## 2   1111.900024  4857900  1123.069946
## 3   1055.800049  3798300  1110.000000
## 4   1080.599976  3448000  1081.709961
## 5   1048.579956  2341700  1081.780029
```

Extract financial data from Internet sources into a DataFrame. There are different sources offering different kind of data. Some sources are:

- Robinhood
- IEX
- World Bank
- OECD
- Eurostat

A complete list of the sources and the usage can be found here: pandas-datareader

Import pandas-datareader

```
from pandas_datareader import data
```

`data.DataReader("stock symbol", "source", "start", "end")`: get financial data of a stock in a certain time period.

Robinhood get data

```
ford = data.DataReader("F", "robinhood", "1/1/2017", "1/31/2018")
print(ford.head()[["close_price", "volume"]])

##                     close_price    volume
## symbol begins_at
## F      2017-10-09     11.575500  28924795
##        2017-10-10     11.622400  40586234
##        2017-10-11     11.613000  34953438
##        2017-10-12     11.369100  45924434
##        2017-10-13     11.303500  44597334
```

Stock code list

Robinhood handle data

```
print(ford.index)
```

```
## MultiIndex(levels=[[F], [2017-01-02 00:00:00, 2017-01-03...
## names=[Symbol, Date])
```

```
print(ford.loc["F", "1/26/2018"])
```

```
## close_price      11.063900
## high_price       11.111400
## interpolated         False
## low_price        10.921500
## open_price       11.007000
## session                reg
## volume            52496001
## Name: (F, 2018-01-26 00:00:00), dtype: object
```

DataFrame index

Index of the DataFrame is different at different sources. Always check `DataFrame.index`!

IEX

```
sap = data.DataReader("SAP", "iex", "1/1/2017", "1/31/2018")
print(sap[25:27])

##                open     high      low    close  volume
## date
## 2017-02-08  89.5382  90.0263  89.4405  89.6065  653804
## 2017-02-09  89.7139  89.9738  89.5284  89.5284  548787

print(sap.loc["2017-02-08"])

## open          89.5382
## high          90.0263
## low           89.4405
## close         89.6065
## volume    653804.0000
## Name: 2017-02-08, dtype: float64
```

Eurostat

```
population = data.DataReader("tps00001", "eurostat", "1/1/2007",
"1/1/2018")
```

```
print(population.columns)
```

```
## MultiIndex(levels=[[Population on 1 January - total], [Albania,
## Andorra, Armenia, Austria, Azerbaijan, Belarus, Belgium, ...
```

```
print(population["Population on 1 January - total", "France"][0:5])
```

```
## FREQ             Annual
## TIME_PERIOD
## 2007-01-01   63645065.0
## 2008-01-01   64007193.0
## 2009-01-01   64350226.0
## 2010-01-01   64658856.0
## 2011-01-01   64978721.0
```

Eurostat Database

Website used for the example: Econometrics

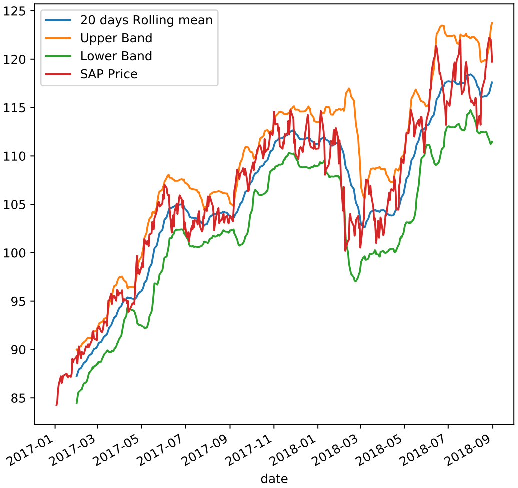
Beautiful Soup

```
from bs4 import BeautifulSoup
import requests
url = "www.uni-goettingen.de/de/applied-econometrics/412565.html"
r = requests.get("https://" + url)
d = r.text
soup = BeautifulSoup(d, "lxml")

print(soup.title)

## <title>Applied Econometrics - Georg-August-... ...</title>
```

Reading data from HTML in detail exceeds the content of this course. If you are interested in this kind of importing data, you can find detailed information on *Beautiful Soup* here.

Bollinger

```
sap = data.DataReader("SAP", "iex", "1/1/2017", "8/31/2018")
sap.index = pd.to_datetime(sap.index)
boll = sap["close"].rolling(window=20, center=False).mean()
std = sap["close"].rolling(window=20, center=False).std()
upp = boll + std * 2
low = boll - std * 2
fig = plt.figure()
ax = fig.add_subplot(1, 1, 1)
boll.plot(ax=ax, label="20 days Rolling mean")
upp.plot(ax=ax, label="Upper Band")
low.plot(ax=ax, label="Lower Band")
sap["close"].plot(ax=ax, label="SAP Price")
ax.legend(loc="best")
fig.savefig("out/boll.pdf")
```


20 days Rolling mean
Upper Band
Lower Band
SAP Price
125
120
115
110
105
100
95
90
85
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
2018-09
date

Visual illustrations

Visual illustrations

- Matplotlib

The package `matplotlib` is a free software library for python including the following functions:

- Image plot, Contour plot, Scatter plot, Polar plot, Line plot, 3-D plot,
- Variety of hardcopy formats,
- Works in Python scripts, the Python and IPython shell and the jupyter notebook,
- Interactive environments.

Usage of matplotlib

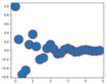
matplotlib has a vast number of functions and options, which is hard to remember. But for almost every task there is an example you can take code from. A great source of information is the examples gallery on the `matplotlib` homepage. Also note the Best practice Quick Start Guide

Gallery

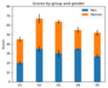
This gallery contains examples of the many things you can do with Matplotlib. Click on any image to see the full image and source code.

For longer tutorials, see our tutorials page. You can also find external resources and a FAQ in our user guide.

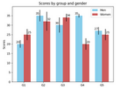
Lines, bars and markers



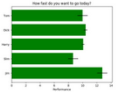
Arctest



Stacked Bar Graph



Barchart

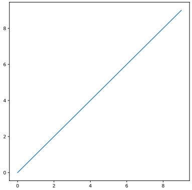


Horizontal bar chart

`plt.plot(array)`: plot the values of a list, the X-axis has by default the range (0, 1, ..., n).

Import matplotlib and simple example

```
import matplotlib.pyplot as plt
import numpy as np
plt.plot(np.arange(10))
plt.savefig("out/list.pdf")
```



Visual illustrations

- Figures and subplots

Plots in `matplotlib` reside in a *Figure object*:

`plt.figure(figsize)` creates new Figure object with multiple options.

`plt.gcf()`: reference of the active figure.

Create Figures

```
fig = plt.figure(figsize=(16, 8))
print(plt.gcf())

## Figure(1600x800)
```

- A Figure object can be considered as an empty window,
- The Figure object has a number of options, such as the size or the aspect ratio,
- You cannot make a plot in a blank figure. There has to be a subplot in the Figure object.

`plt.savefig("filename")`: Saving active figure to file.
Available file formats are among others:

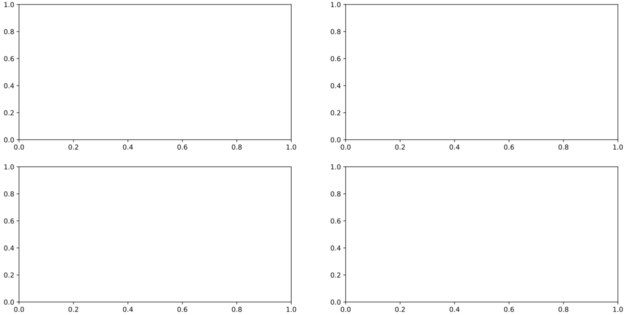
Filename extension	Description
.png	Portable Network Graphics
.pdf	Portable Document Format
.svg	Scalable Vector Graphics
.jpeg	JPEG File Interchange Format
.jpg	JPEG File Interchange Format
.ps	PostScript
.raw	Raw image format

`fig.add_subplot()`: adds subplot to the Figure fig.
Example: `fig.add_subplot(2, 2, 1)` creates four subplots and selects the first.

Adding subplots

```
ax1 = fig.add_subplot(2, 2, 1)
ax2 = fig.add_subplot(2, 2, 2)
ax3 = fig.add_subplot(2, 2, 3)
ax4 = fig.add_subplot(2, 2, 4)
fig.savefig("out/subplots.pdf")
```

- The Figure object is filled with subplots in which the plots reside,
- Using the `plt.plot()` command without creating a subplot in advance, matplotlib will create a Figure object and a subplot automatically,
- The Figure object and its subplots can be created in one line.

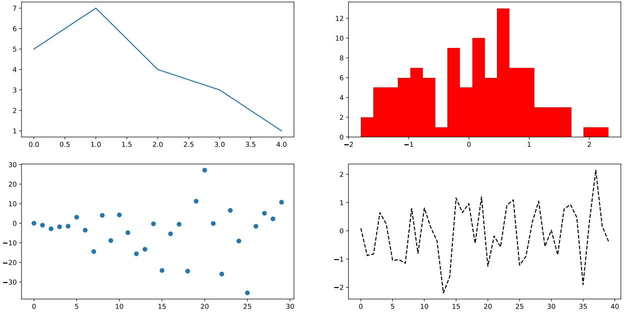

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Filling subplots with content

```
from numpy.random import randn
ax1.plot([5, 7, 4, 3, 1])
ax2.hist(randn(100), bins=20, color="r")
ax3.scatter(np.arange(30), np.arange(30)*randn(30))
ax4.plot(randn(40), "k--")
fig.savefig("out/content.pdf")
```

- The subplots in one Figure object can be filled with different plot types,
- Using only `plt.plot()` matplotlib draws the plot in the last Figure object and last subplot selected.

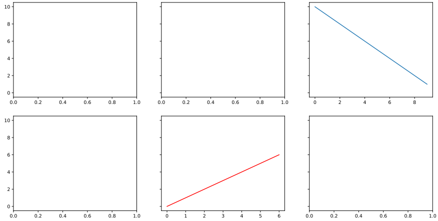
Subplots


1 2 3 4 5 6 7
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0 2 4 6 8 10 12
−2 −1 0 1 2
−30 −20 −10 0 10 20 30
0 5 10 15 20 25 30
−2 −1 0 1 2
0 5 10 15 20 25 30 35 40

`plt.subplots(nrows, ncols, sharex, sharey)`: creates figure and subplots in one line. If `sharex` or `sharey` are `True`, all subplots share the same X- or Y-ticks.

Standard creation

```
fig, axes = plt.subplots(2, 3, figsize=(16, 8), sharey=True)
axes[1, 1].plot(np.arange(7), color="r")
axes[0, 2].plot(np.arange(10, 0, -1))
fig.savefig("out/standard.pdf")
```



Visual illustrations

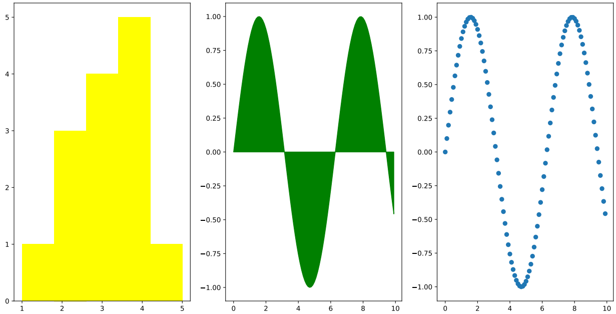
▶ Plot types and styles

`ax.scatter(x, y)`: create a scatter plot of `x` vs `y`.
`ax.hist(x, bins)`: create a histogram.
`ax.fill_between(x, y, a)`: create a plot of `x` vs `y` and fills plot between `a` and `y`.

Types

```
fig, ax = plt.subplots(1, 3, figsize=(16, 8))
ax[0].hist([1, 2, 3, 4, 5, 4, 3, 2, 3, 4, 2, 3, 4, 4], bins=5,
color="yellow")
x = np.arange(0, 10, 0.1)
y = np.sin(x)
ax[1].fill_between(x, y, 0, color="green")
ax[2].scatter(x, y)
fig.savefig("out/types.pdf")
```

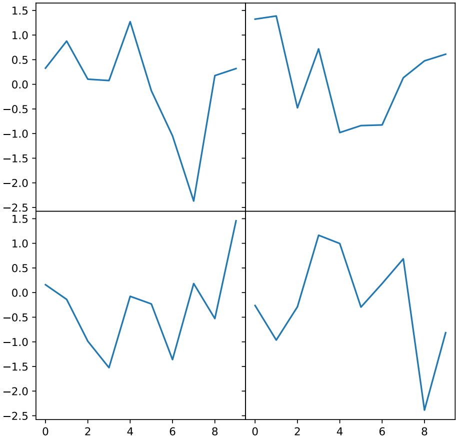
A vast number of plot types can be found in the examples gallery.


5
4
3
2
1
0
1 2 3 4 5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0 2 4 6 8 10
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0 2 4 6 8 10

`plt.subplots_adjust(left, bottom, ..., hspace)`: set the space between the subplots. `wspace` and `hspace` control the percentage of the figure width and figure height, respectively, to use as spacing between subplots.

Adjust spacing

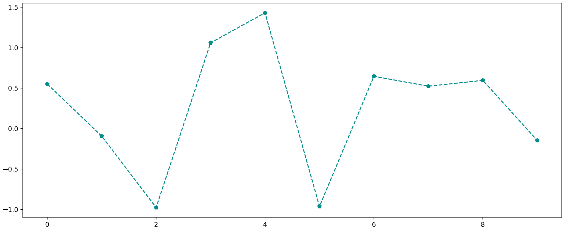
```
fig, axes = plt.subplots(2, 2, sharex=True, sharey=True)
for i in range(2):
    for j in range(2):
        axes[i][j].plot(randn(10))
plt.subplots_adjust(wspace=0, hspace=0)
fig.savefig("out/spacing.pdf")
```


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0
2
4
6
8

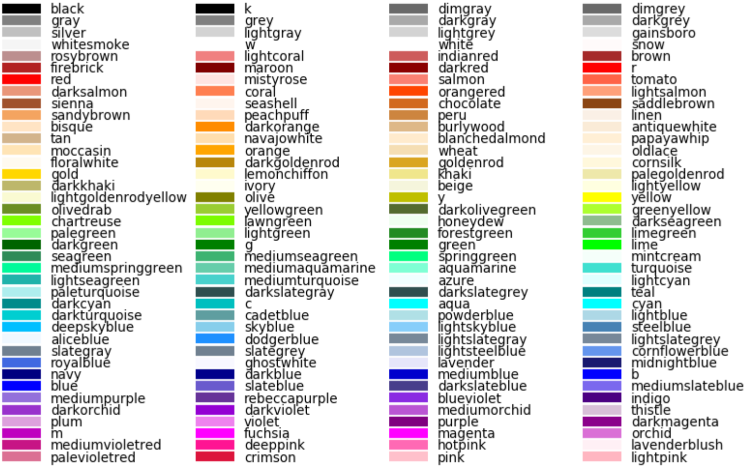
`ax.plot(data, linestyle, color, marker)`: set data and styles of subplot `ax`.

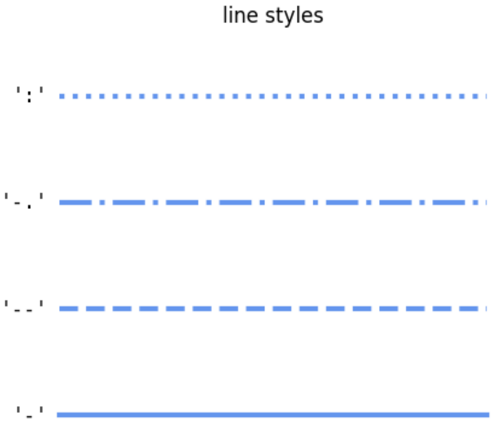
Styles

```
fig, ax = plt.subplots(1, figsize=(15, 6))
ax.plot(randn(10), linestyle="--", color="darkcyan", marker="p")
fig.savefig("out/style.pdf")
```



Plot colors


black
gray
silver
whitesmoke
rosybrown
firebrick
red
darksalmon
sienna
sandybrown
bisque
tan
moccasin
floralwhite
gold
darkkhaki
lightgoldenrodyellow
olivedrab
chartreuse
palegreen
darkgreen
seagreen
mediumspringgreen
lightseagreen
paleturquoise
darkcyan
darkturquoise
deepskyblue
aliceblue
slategray
royalblue
navy
blue
mediumpurple
darkorchid
plum
m
mediumvioletred
palevioletred
k
grey
lightgray
w
lightcoral
maroon
mistyrose
coral
seashell
peachpuff
darkorange
navajowhite
orange
darkgoldenrod
lemonchiffon
ivory
olive
yellowgreen
lawngreen
lightgreen
g
mediumseagreen
mediumaquamarine
mediumturquoise
darkslategray
c
cadetblue
skyblue
dodgerblue
slategrey
ghostwhite
darkblue
slateblue
rebeccapurple
darkviolet
violet
fuchsia
deeppink
crimson
dimgray
darkgray
lightgrey
white
indianred
darkred
salmon
orangered
chocolate
peru
burlywood
blanchedalmond
wheat
goldenrod
khaki
beige
y
darkolivegreen
honeydew
forestgreen
green
springgreen
aquamarine
azure
darkslategrey
aqua
powderblue
lightskyblue
lightslategray
lightsteelblue
lavender
mediumblue
darkslateblue
blueviolet
mediumorchid
purple
magenta
hotpink
pink
dimgrey
darkgrey
gainsboro
snow
brown
r
tomato
lightsalmon
saddlebrown
linen
antiquewhite
papayawhip
oldlace
cornsilk
palegoldenrod
lightyellow
yellow
greenyellow
darkseagreen
limegreen
lime
mintcream
turquoise
lightcyan
teal
cyan
lightblue
steelblue
lightslategrey
cornflowerblue
midnightblue
b
mediumslateblue
indigo
thistle
darkmagenta
orchid
lavenderblush
lightpink


line styles
':'
'-.'
'--'
'-'

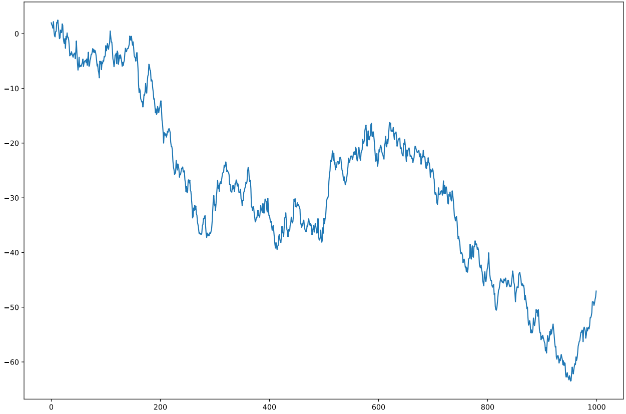
Marker	Description
"."	point
","	pixel
"o"	circle
"v"	triangle_down
"8"	octagon
"s"	square
"p"	pentagon
"P"	plus (filled)
"*"	star
"h"	hexagon1
"H"	hexagon2
"+"	plus
"x"	x
"X"	x (filled)
"D"	diamond

`ax.set_xticks()`: set list of X-ticks, alalogous for Y-axis.
`ax.set_xlabel()`: set the X-label.
`ax.set_title()`: set the subplot title.

Ticks and labels - default

```
fig, ax = plt.subplots(1, figsize=(15, 10))
ax.plot(randn(1000).cumsum())
fig.savefig("out/withoutlabls.pdf")
```

- Here a Figure object and a subplot were created and filled with a plot,
- By default matplotlib places the ticks evenly distributed along the data range. Individual ticks can be set as follows,
- By default there is no axis label or title.

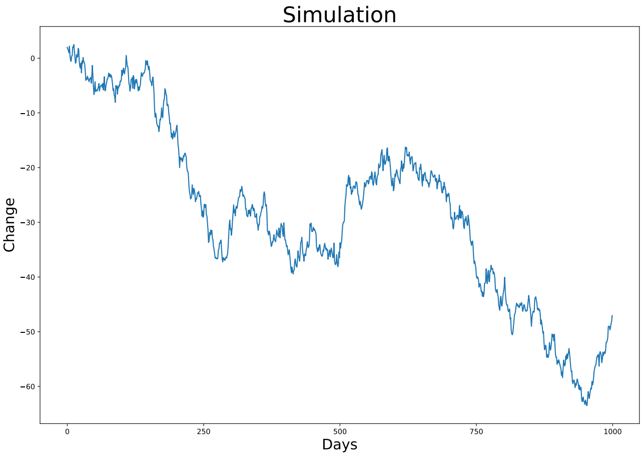


0
−10
−20
−30
−40
−50
−60
0
200
400
600
800
1000

Set ticks and labels

```
ax.set_xticks([0, 250, 500, 750, 1000])
ax.set_xlabel("Days", fontsize=20)
ax.set_ylabel("Change", fontsize=20)
ax.set_title("Simulation", fontsize=30)
fig.savefig("out/labels.pdf")
```

- The individual ticks are given as a list to `ax.set_xticks()`,
- The label and titel can be set to an individual size using the argument `fontsize`.



Simulation
Change
Days
0
−10
−20
−30
−40
−50
−60
0
250
500
750
1000

Legends

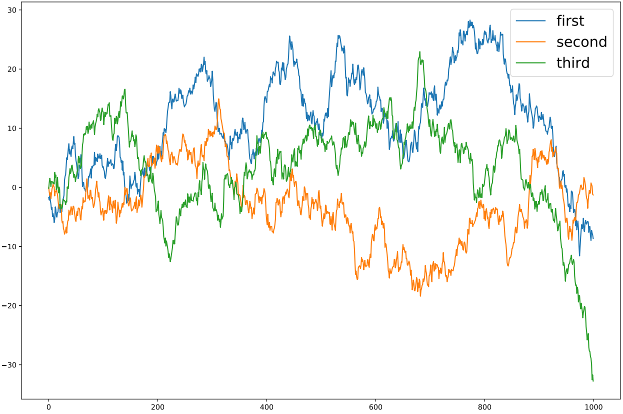
Using multiple plots in one subplot one needs a legend.
`ax.legend(loc)`: showing the legend at location `loc`.
Some options: `"best"`, `"upper right"`, `"center left"`, ...

Set legend

```
fig = plt.figure(figsize=(15, 10))
ax = fig.add_subplot(1, 1, 1)
ax.plot(randn(1000).cumsum(), label="first")
ax.plot(randn(1000).cumsum(), label="second")
ax.plot(randn(1000).cumsum(), label="third")
ax.legend(loc="best", fontsize=20)
fig.savefig("out/legend.pdf")
```

- The legend displays the label and the color of the associated plot,
- Using the option `"best"` the legend will placed in a corner where is does not interfere the plots.

Legends

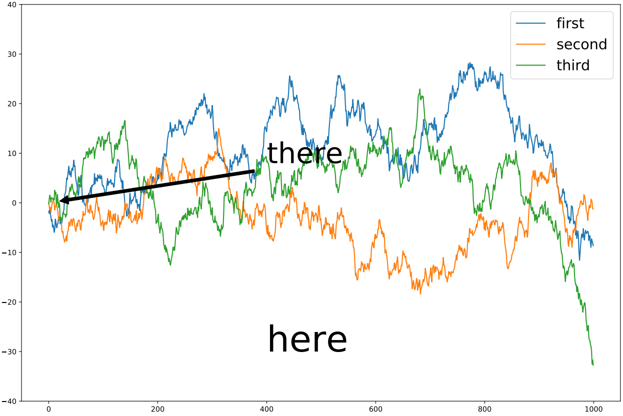


`ax.text(x, y, "text", fontsize)`: insert text into a subplot.
`ax.annotate("text", xy, xytext, arrwoprops)`: insert arrow with annotations.

Annotations

```
ax.text(400, -30, "here", fontsize=50)
ax.annotate("there",
            fontsize=40,
            xy=(0, 0),
            xytext=(400, 8),
            arrowprops=dict(facecolor="black",
                            shrink=0.05))
ax.set_yticks([-40, -30, -20, -10, 0, 10, 20, 30, 40])
fig.savefig("out/arrow.pdf")
```

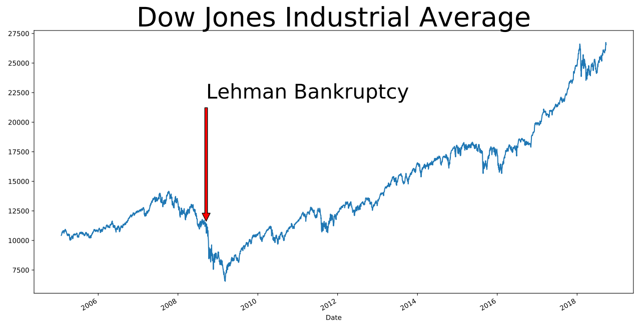
- Using `ax.annotate()` the arrow head points at `xy` and the bottom left corner of the text will be placed at `xytext`.


first
second
third
there
here
40
30
20
10
0
−10
−20
−30
−40
0
200
400
600
800
1000

Annotation Lehman

```
import pandas as pd
from datetime import datetime
date = datetime(2008, 9, 15)
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
dow = pd.read_csv("data/dji.csv", index_col=0, parse_dates=True)
close = dow["Close"]
close.plot(ax=ax)
ax.annotate("Lehman Bankruptcy",
            fontsize=30,
            xy=(date, close.loc[date] + 400),
            xytext=(date, 22000),
            arrowprops=dict(facecolor="red",
                            shrink=0.03))
ax.set_title("Dow Jones Industrial Average", size=40)
fig.savefig("out/lehman.pdf")
```

Annotations

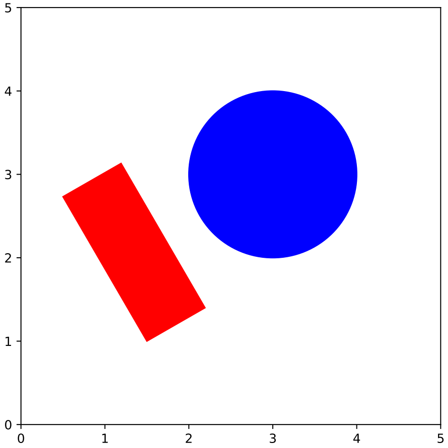


`plt.Rectangle((x, y), width, height, angle)`: create a rectangle
`plt.Circle((x,y), radius)`: create a circle.

Drawing

```
fig = plt.figure(figsize=(6, 6))
ax = fig.add_subplot(1, 1, 1)
ax.set_xticks([0, 1, 2, 3, 4, 5])
ax.set_yticks([0, 1, 2, 3, 4, 5])
rectangle = plt.Rectangle((1.5, 1),
                          width=0.8, height=2,
                          color="red", angle=30)
circ = plt.Circle((3, 3),
                  radius=1, color="blue")
ax.add_patch(rectangle)
ax.add_patch(circ)
fig.savefig("out/draw.pdf")
```

A list of all available patches can be found here: matplotlib-patches



0
1
2
3
4
5
0
1
2
3
4
5

Step 1

Create a Figure object and subplots

Best practice Step 1

```
fig, ax = plt.subplots(1, 1, figsize=(16, 8))
```

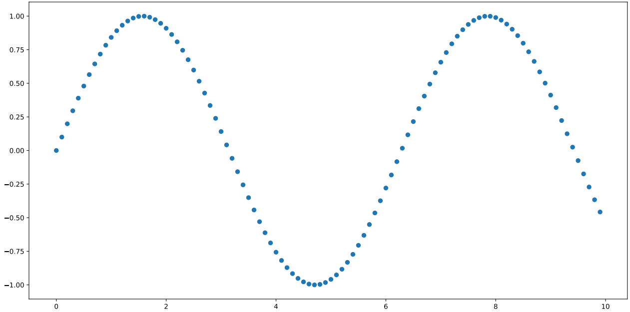
Step 2

Plot data using different plot types

An overview of plot types can be found in the examples gallery.

Best practice Step 2

```
x = np.arange(0, 10, 0.1)
y = np.sin(x)
ax.scatter(x, y)
```


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10

Step 3
Set colors, markers and line styles

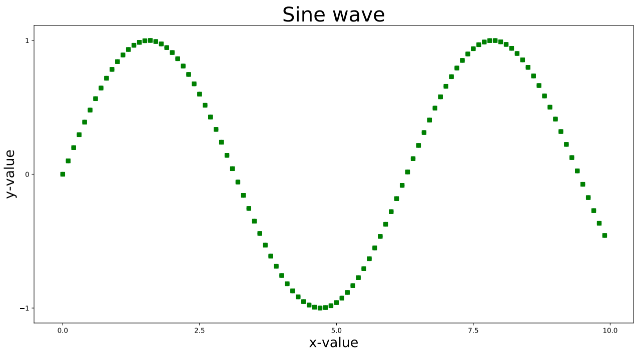
Best practice Step 3

```
ax.scatter(x, y, color="green", marker="s")
```

Step 4
Set title, axis labels and ticks

Best practice Step 4

```
ax.set_title("Sine wave", fontsize=30)
ax.set_xticks([0, 2.5, 5, 7.5, 10])
ax.set_yticks([-1, 0, 1])
ax.set_ylabel("y-value", fontsize=20)
ax.set_xlabel("x-value", fontsize=20)
```


Sine wave
y-value
1
0
−1
0.0
2.5
5.0
7.5
10.0
x-value

Step 5
Set labels

Best practice Step 5

```
ax.scatter(x, y, color="green", marker="s", label="Sine")
```

Step 6
Set legend (if you add another plot to existing subfigure)

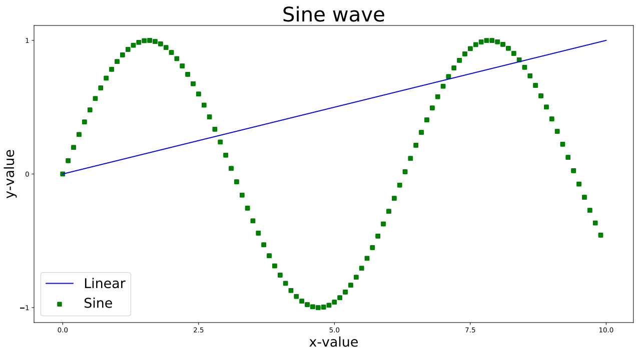
Best practice Step 6

```
ax.plot(np.arange(11)/10, color="blue", linestyle="-",
label="Linear")
ax.legend(fontsize=20)
```

Step 7
Save plot to file

Best practice Step 7

```
fig.savefig("out/sinewave.pdf")
```


Sine wave
1
0
−1
y-value
0.0
2.5
5.0
7.5
10.0
x-value
Linear
Sine

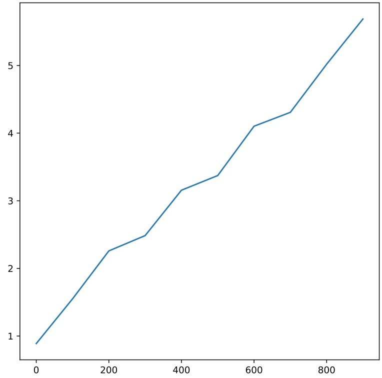
Visual illustrations

- Pandas visualization

`DataFrame/Series.plot()`: plot a DataFrame or a Series.

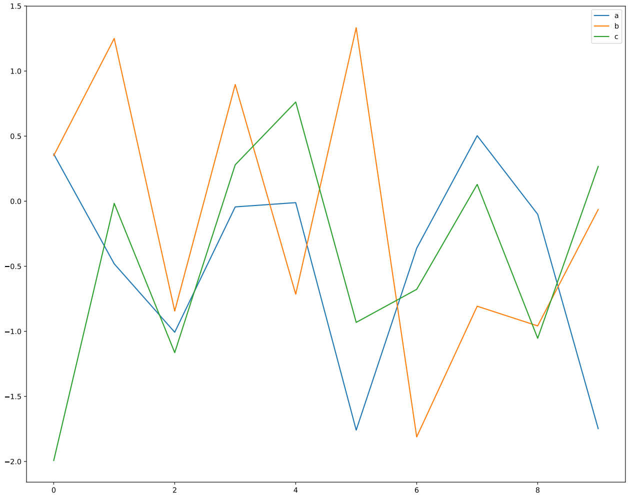
Simple line plot

```
plt.close("all")
p = pd.Series(np.random.rand(10).cumsum(), index=np.arange(0, 1000,
100))
print(p)

## 0      0.888442
## 100    1.549929
## 200    2.258732
## 300    2.485168
## 400    3.156098
## 500    3.373227
## 600    4.102376
## 700    4.307634
## 800    5.019096
## 900    5.687669
## dtype: float64

p.plot()
plt.savefig("out/line.pdf")
```


5
4
3
2
1
0
200
400
600
800

Line plots

```
df = pd.DataFrame(np.random.randn(10, 3), index=np.arange(10),
                  columns=["a", "b", "c"])
print(df)

##           a         b         c
## 0  0.362041  0.350474 -1.992641
## 1 -0.481396  1.250534 -0.017076
## 2 -1.007017 -0.843875 -1.163215
## 3 -0.043806  0.896435  0.279640
## 4 -0.011092 -0.714289  0.762072
## 5 -1.758891  1.332606 -0.931393
## 6 -0.361416 -1.811150 -0.677346
## 7  0.503350 -0.806999  0.129074
## 8 -0.100652 -0.958269 -1.053158
## 9 -1.747851 -0.064166  0.267087

df.plot(figsize=(15, 12))
plt.savefig("out/line2.pdf")
```


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
2
4
6
8
a
b
c

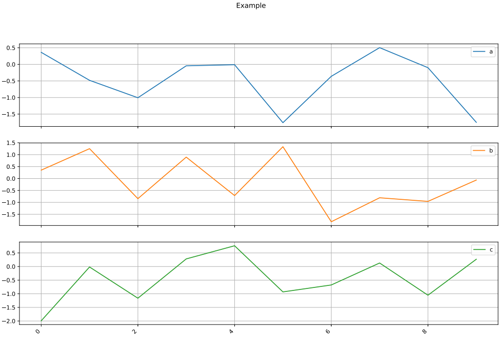
The plot method applied to a DataFrame plots each column as a different line and shows the legend automatically. Plotting DataFrames, there are serveral arguments to change the style of the plot:

Argument	Description
kind	"line", "bar", etc
logy	logarithmic scale on Y-axis
use_index	If True, use index for tick labels
rot	Rotation of tick labels
xticks	Values for x ticks
yticks	Values for y ticks
grid	Set grid True or False
xlim	X-axis limits
ylim	Y-axis limits
subplots	Plot each DataFrame column in a new subplot

Table: pandas plot arguments

Separated line plots

```
df.plot(grid=True, rot=45, subplots=True, title="Example",
        figsize=(15, 10))
plt.savefig("out/pandas.pdf")
```



`DataFrame.plot(ax = subplot)`: plot DataFrame into an existing subplot.

Standard creation

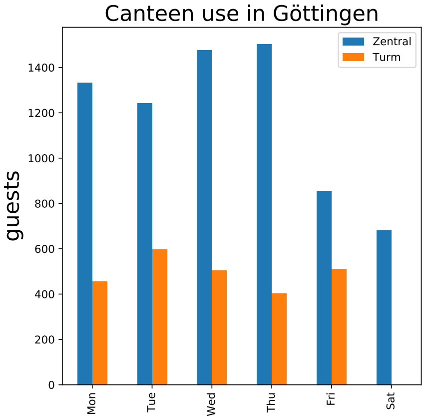
```
fig = plt.figure(figsize=(6, 6))
ax = fig.add_subplot(1, 1, 1)
guests = np.array([[1334, 456], [1243, 597], [1477, 505],
                   [1502, 404], [854, 512], [682, 0]])
canteen = pd.DataFrame(guests,
                       index=["Mon", "Tue", "Wed",
                              "Thu", "Fri", "Sat"],
                       columns=["Zentral", "Turm"])
print(canteen)

##      Zentral  Turm
## Mon     1334   456
## Tue     1243   597
## Wed     1477   505
## Thu     1502   404
## Fri      854   512
## Sat      682     0
```

Bar plot

```
canteen.plot(ax=ax, kind="bar")
ax.set_ylabel("guests", fontsize=20)
ax.set_title("Canteen use in Göttingen", fontsize=20)
fig.savefig("out/canteen.pdf")
```

- The bar plot resides in the subplot `ax`,
- The label and title are set as shown before without using pandas.

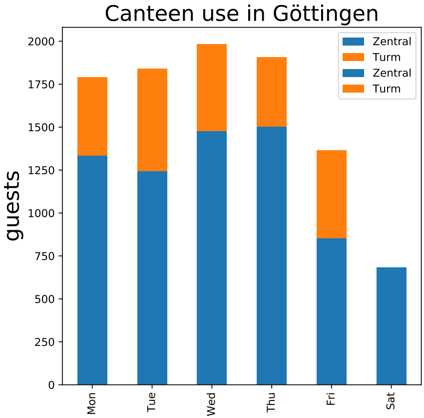
Bar plot


Canteen use in Göttingen
Zentral
Turm
guests
0
200
400
600
800
1000
1200
1400
Mon
Tue
Wed
Thu
Fri
Sat

Bar plot - stacked

```
canteen.plot(ax=ax, kind="bar", stacked=True)
ax.set_ylabel("guests", fontsize=20)
ax.set_title("Canteen use in Göttingen", fontsize=20)
fig.savefig("out/canteenstacked.pdf")
```

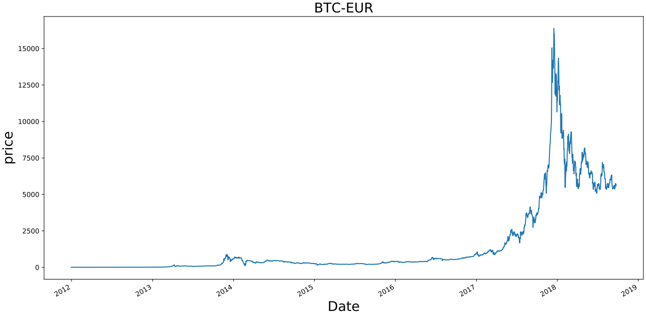
Bar plot


Canteen use in Göttingen
guests
2000
1750
1500
1250
1000
750
500
250
0
Mon
Tue
Wed
Thu
Fri
Sat
Zentral
Turm
Zentral
Turm

BTC chart

```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel("price", fontsize=20)
ax.set_xlabel("Date", fontsize=20)
BTC = pd.read_csv("data/btc-eur.csv", index_col=0, parse_dates=True)
BTCclose = BTC["Close"]
BTCclose.plot(ax=ax)
ax.set_title("BTC-EUR", fontsize=20)
fig.savefig("out/btc.pdf")
```

Plot financial data

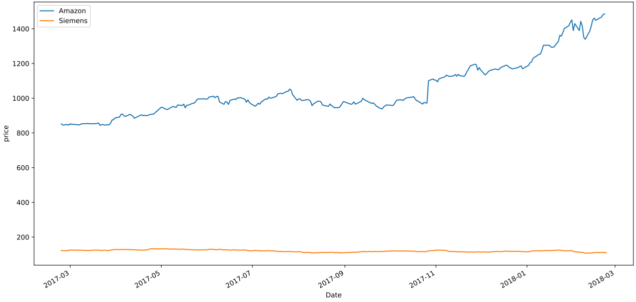


Compare - bad illustration

```
amazon = pd.read_csv("data/amzn.csv", index_col=0,
                     parse_dates=True)["Close"]
siemens = pd.read_csv("data/sie.de.csv", index_col=0,
                      parse_dates=True)["Close"]
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel("price")
amazon.plot(ax=ax, label="Amazon")
siemens.plot(ax=ax, label="Siemens")
ax.legend(loc="best")
fig.savefig("out/compare.pdf")
```

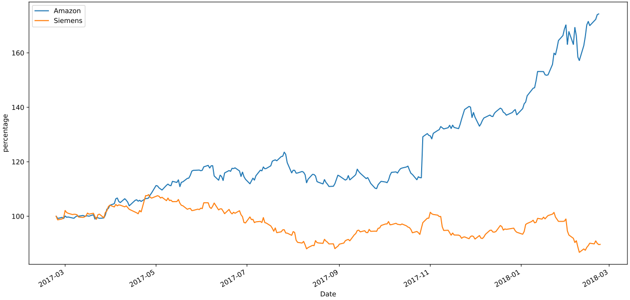
- In this illustration you can hardly compare the trend of the two stocks,
- Using pandas you can standardize both dataframes in one line.

Plot financial data



Compare - good illustration

```
amazon = amazon/amazon[0] * 100
siemens = siemens/siemens[0] * 100
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel("percentage")
amazon.plot(ax=ax, label="Amazon")
siemens.plot(ax=ax, label="Siemens")
ax.legend(loc="best")
fig.savefig("out/comparenew.pdf")
```


Amazon
Siemens
percentage
100
120
140
160
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

Applications

Applications

▶ Time series

Data types for date and time are included in the Python standard library.

Datetime creation

```
from datetime import datetime
now = datetime.now()
print(now)

## 2018-10-08 22:53:27.197198

print(now.day)

## 8

print(now.hour)

## 22
```

From datetime you can get the attributes `year`, `month`, `day`, `hour`, `second`.

`datetime(year, month, day, hour, minute, second)`: set time and date.

Datetime representation

```
holiday = datetime(2018, 12, 24, 8, 30)
print(holiday)

## 2018-12-24 08:30:00

exam = datetime(2018, 11, 9, 10)
print("The exam will be on the " + "{:%Y-%m-%d}".format(exam))

## The exam will be on the 2018-11-09
```

`timedelta(days, seconds)`: represent difference between two date-time objects.

Datetime difference

```
from datetime import timedelta
delta = exam - now
print(delta)

## 31 days, 11:06:32.802802

print("The exam will take place in " + str(delta.days) + " days.")

## The exam will take place in 31 days.

print(now)

## 2018-10-08 22:53:27.197198

print(now + timedelta(10, 120))

## 2018-10-18 22:55:27.197198
```

Convert string and datetime

`datetime.strftime("format")`: convert datetime object into string.
`datetime.strptime(datestring, "format")`: convert date as a string into a datetime object.

Convert Datetime

```
stamp = datetime(2018, 4, 12)
print(stamp)

## 2018-04-12 00:00:00

print("German date format: " + stamp.strftime("%d.%m.%Y"))

## German date format: 12.04.2018

val = "2018-5-5"
d = datetime.strptime(val, "%Y-%m-%d")
print(d)

## 2018-05-05 00:00:00
```

Converting examples

```
val = "31.01.2012"
d = datetime.strptime(val, "%d.%m.%Y")
print(d)

## 2012-01-31 00:00:00

print(now.strftime("Today is %A and we are in week %W of the year
%Y."))

## Today is Monday and we are in week 41 of the year 2018.

print(now.strftime("%c"))

## Mon 08 Oct 2018 10:53:27 PM
```

Type	Description
%Y	4-digit year
%m	2-digit month [01, 12]
%d	2-digit day [01, 31]
%H	Hour (24-hour clock) [00, 23]
%I	Hour (12-hour clock) [01, 12]
%M	2-digit minute [00, 59]
%S	Second [00, 61]
%W	Week number of the year [00, 53]
%F	Shortcut for %Y-%m-%d

Type	Description
%a	Abbreviated weekday name
%A	Full weekday name
%b	Abbreviated month name
%B	Full month name
%c	Full date and time
%x	Locale-appropriate formatted date

`pd.date_range(start, end, freq)`: generate a date range.

Date ranges

```
import pandas as pd
index = pd.date_range("2018-01-01", now)
print(index[0:2])
print(index[15:16])
index = pd.date_range("2018-01-01", now, freq="M")
print(index[0:2])
```

```
## DatetimeIndex(['2018-01-01', '2...ype='datetime64[ns]', freq='D')
## DatetimeIndex(['2018-01-16'], dtype='datetime64[ns]', freq='D')
## DatetimeIndex(['2018-01-31', '2...ype='datetime64[ns]', freq='M')
```

Alias	**Offset type**
D	Day
B	Business day
H	Hour
T	Minute
S	Second
M	Month end
BM	Business month end
Q-JAN, Q-FEB, ...	Quarter end
A-JAN, A-FEB, ...	Year end
AS-JAN, AS-FEB, ...	Year begin
BA-JAN, BA-FEB, ...	Business year end
BAS-JAN, BAS-FEB, ...	Business year begin

`DataFrame.resample("frequency")`: resample frecency of time series.

Resample date ranges

```
import numpy as np
start = datetime(2016, 1, 1)
ind = pd.date_range(start, now)
numbers = np.arange((now - start).days + 1)
df = pd.DataFrame(numbers, index=ind)
```

```
print(df.head())

##              0
## 2016-01-01   0
## 2016-01-02   1
## 2016-01-03   2
## 2016-01-04   3
## 2016-01-05   4
```

```
print(df.resample("3BM").sum().head())

##                  0
## 2016-01-29     406
## 2016-04-29    6734
## 2016-07-29   15015
## 2016-10-31   24205
## 2017-01-31   32246
```

Applications

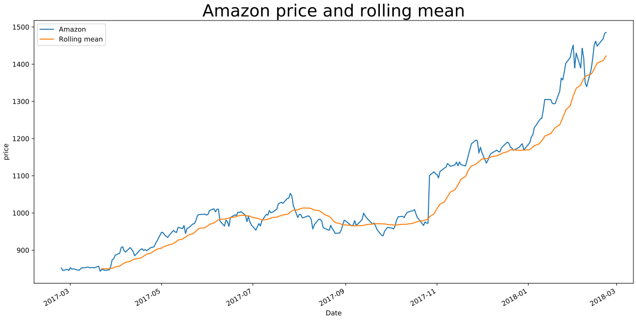
▶ Moving window

`DataFrame.rolling(window)`: conduct rolling window computations.

Rolling mean

```
import matplotlib.pyplot as plt
amazon = pd.read_csv("data/amzn.csv", index_col=0,
                     parse_dates=True)["Adj Close"]
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
ax.set_ylabel("price")
amazon.plot(ax=ax, label="Amazon")
amazon.rolling(window=20).mean().plot(ax=ax, label="Rolling mean")
ax.legend(loc="best")
ax.set_title("Amazon price and rolling mean", fontsize=25)
fig.savefig("out/amzn.pdf")
```

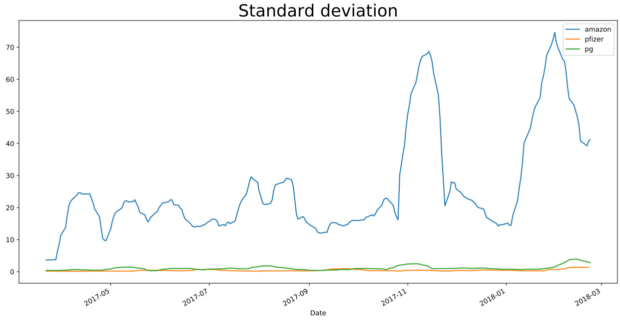
Rolling functions are: `mean()`, `median()`, `sum()`, `var()`, `std()`, `min()`, `max()`.



Amazon price and rolling mean
Amazon
Rolling mean
price
Date
1500
1400
1300
1200
1100
1000
900
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03

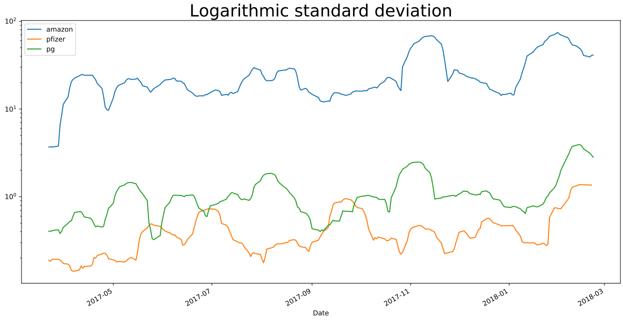
Standard deviation

```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
pfizer = pd.read_csv("data/pfe.csv", index_col=0,
                     parse_dates=True)["Adj Close"]
pg = pd.read_csv("data/pg.csv", index_col=0,
                 parse_dates=True)["Adj Close"]
all = pd.DataFrame(index=amazon.index)
all["amazon"] = pd.DataFrame(amazon)
all["pfizer"] = pd.DataFrame(pfizer)
all["pg"] = pd.DataFrame(pg)
all_std = all.rolling(window=20).std()
all_std.plot(ax=ax)
ax.set_title("Standard deviation", fontsize=25)
fig.savefig("out/std.pdf")
```


Standard deviation
amazon
pfizer
pg
70
60
50
40
30
20
10
0
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

Logarithmic standard deviation

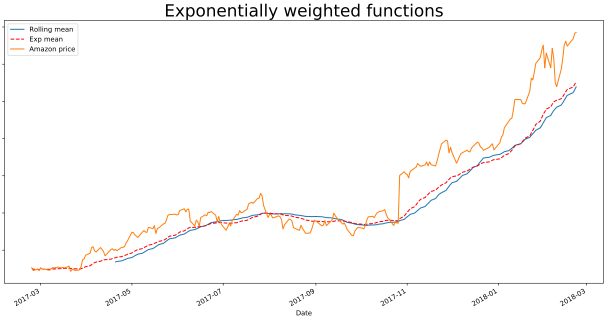
```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
all_std.plot(ax=ax, logy=True)
ax.set_title("Logarithmic standard deviation", fontsize=25)
fig.savefig("out/std_log.pdf")
```


Logarithmic standard deviation
amazon
pfizer
pg
10^2
10^1
10^0
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

`DataFrame.ewm(span)`: compute exponentially weighted rolling window functions.

Exponentially weighted functions

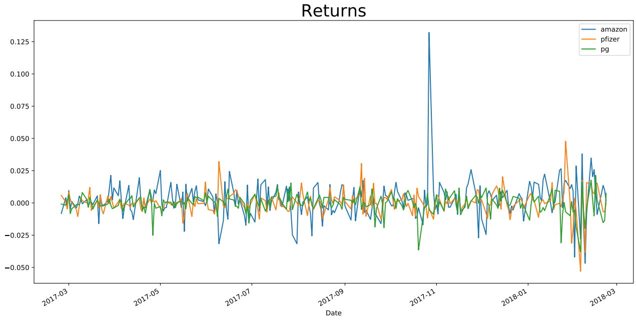
```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
amazon.rolling(window=40).mean().plot(ax=ax, label="Rolling mean")
amazon.ewm(span=40).mean().plot(ax=ax, label="Exp mean",
                                linestyle="--", color="red")
amazon.plot(ax=ax, label="Amazon price")
ax.legend(loc="best")
ax.set_title("Exponentially weighted functions", fontsize=25)
fig.savefig("out/mean.pdf")
```


Exponentially weighted functions
Rolling mean
Exp mean
Amazon price
1500
1400
1300
1200
1100
1000
900
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

`DataFrame.pct_change()`: get the daily percentage change.

Percentage change

```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
returns = all.pct_change()
print(returns.head())

##                amazon    pfizer        pg
## Date
## 2017-02-23        NaN       NaN       NaN
## 2017-02-24 -0.008155  0.005872 -0.000878
## 2017-02-27  0.004023  0.000584 -0.001757
## 2017-02-28 -0.004242 -0.004668  0.001980
## 2017-03-01  0.009514  0.008792  0.006479

returns.plot(ax=ax)
ax.set_title("Returns", fontsize=25)
fig.savefig("out/returns.pdf")
```

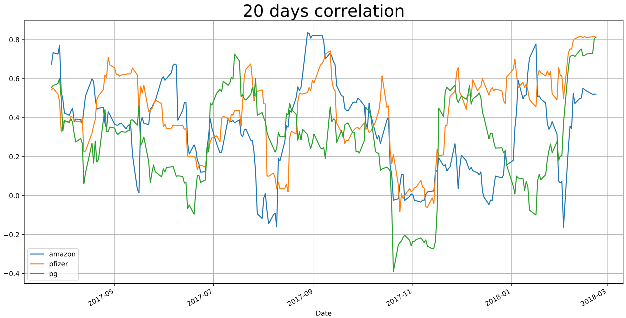

Returns
amazon
pfizer
pg
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

`DataFrame.rolling().corr(benchmark)`: compute correlation between two time series.

Correlation

```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
DJI = pd.read_csv("data/dji.csv", index_col=0,
parse_dates=True)["Adj Close"]
DJI_ret = DJI.pct_change()
corr = returns.rolling(window=20).corr(DJI_ret)
corr.plot(ax=ax)
ax.grid()
ax.set_title("20 days correlation", fontsize=25)
fig.savefig("out/corr.pdf")
```

Binary moving window functions

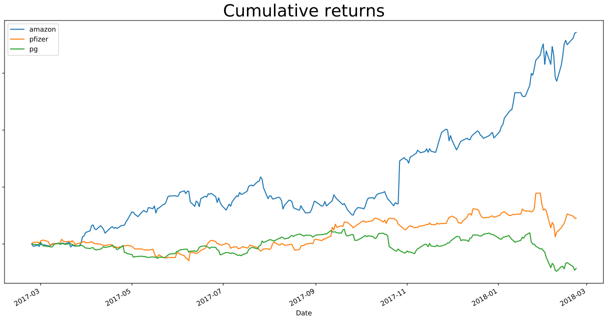


Applications

▶ Financial applications

Returns

```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
ret_index = (1+returns).cumprod()
stocks = ["amazon", "pfizer", "pg"]
for i in stocks:
    ret_index[i][0] = 1
print(ret_index.tail())

##                amazon    pfizer        pg
## Date
## 2018-02-15  1.715298  1.088693  0.932322
## 2018-02-16  1.699961  1.105461  0.934471
## 2018-02-20  1.723031  1.097840  0.920217
## 2018-02-21  1.740128  1.090218  0.907772
## 2018-02-22  1.742968  1.090218  0.914560

ret_index.plot(ax=ax)
ax.set_title("Cumulative returns", fontsize=25)
fig.savefig("out/cumret.pdf")
```


Cumulative returns
amazon
pfizer
pg
1.6
1.4
1.2
1.0
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

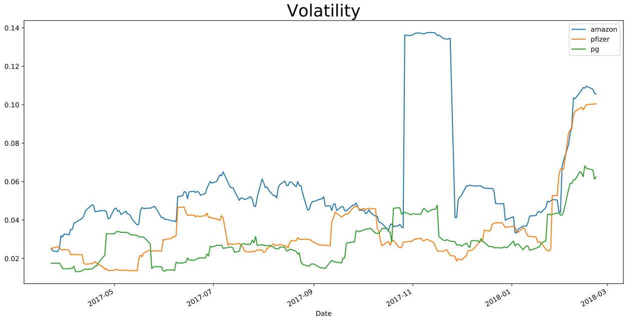
Monthly returns

```
returns_m = ret_index.resample("BM").last().pct_change()
print(returns_m.head())
```

```
##                 amazon    pfizer        pg
## Date
## 2017-02-28         NaN       NaN       NaN
## 2017-03-31   0.049110  0.002638 -0.013396
## 2017-04-28   0.043371 -0.008477 -0.020604
## 2017-05-31   0.075276 -0.028124  0.008703
## 2017-06-30  -0.026764  0.028790 -0.010671
```

Volatility

```
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
vola = returns.rolling(window=20).std() * np.sqrt(20)
vola.plot(ax=ax)
ax.set_title("Volatility", fontsize=25)
fig.savefig("out/vola.pdf")
```


Volatility
amazon
pfizer
pg
0.14
0.12
0.10
0.08
0.06
0.04
0.02
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Date

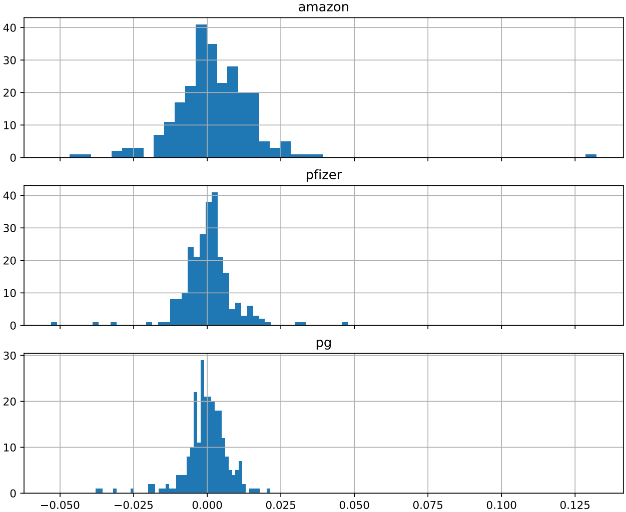
`DataFrame.describe()`: show summarized analysis.

Describe

```
print(all.describe())

##              amazon      pfizer          pg
## count    252.000000  251.000000  252.000000
## mean    1044.521903   33.892665   87.934304
## std      158.041844    1.694680    2.728659
## min      843.200012   30.872143   79.919998
## 25%      953.567474   32.593733   86.241475
## 50%      988.680023   33.147469   87.863598
## 75%     1136.952484   35.331834   90.363035
## max     1485.339966   38.661823   92.988976
```

Histogram

```
fig, ax = plt.subplots(3, 1, figsize=(10, 8), sharex=True)
for i in range(3):
    ax[i].set_title(stocks[i])
    returns[stocks[i]].hist(ax=ax[i], bins=50)
fig.savefig("out/return_hist.pdf")
```


amazon
40
30
20
10
0
pfizer
40
30
20
10
0
pg
30
20
10
0
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
0.125

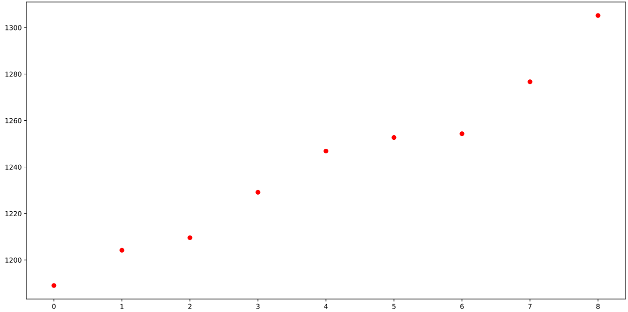
Using the statsmodels module to determine regressions:

`DataFrame.tolist()`: return a list containing the DataFrame values.

`sm.OLS(X, Y).fit()`: get OLS fit of data (`X`, `Y`).

Regression data

```
import statsmodels.api as sm
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
Y = np.array(amazon.loc["2018-1-1":"2018-1-15"].tolist())
X = np.arange(len(Y))
ax.scatter(x=X, y=Y, marker="o", color="red")
fig.savefig("out/reg_data.pdf")
```

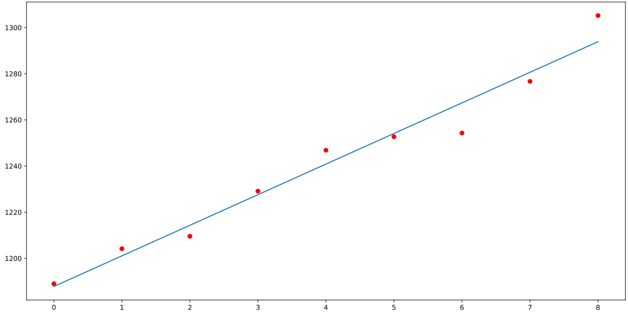

1300
1280
1260
1240
1220
1200
0
1
2
3
4
5
6
7
8

Regression

```
X_reg = sm.add_constant(X)
res = sm.OLS(Y, X_reg).fit()
b, a = res.params
ax.plot(X, a*X + b)
fig.savefig("out/ols.pdf")
```

Summary of OLS regression. To print in python use `res.summary()`.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.965
Model:                            OLS   Adj. R-squared:                  0.959
Method:                 Least Squares   F-statistic:                     190.2
Date:                Mo, 19 Mär 2018   Prob (F-statistic):            2.49e-06
Time:                        15:21:30   Log-Likelihood:                -29.706
No. Observations:                   9   AIC:                             63.41
Df Residuals:                       7   BIC:                             63.81
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1187.8418      4.575    259.617      0.000    1177.023    1198.661
x1            13.2540      0.961     13.792      0.000      10.982      15.526
==============================================================================
Omnibus:                        0.788   Durbin-Watson:                   1.627
Prob(Omnibus):                  0.674   Jarque-Bera (JB):                0.117
Skew:                          -0.268   Prob(JB):                        0.943
Kurtosis:                       2.841   Cond. No.                         9.06
==============================================================================
```


1300
1280
1260
1240
1220
1200
0
1
2
3
4
5
6
7
8

The Newton-Raphson method is an algorithm for finding successively better approximations to the roots of real-valued functions.

Let $F : \mathbb{R}^k \to \mathbb{R}^k$ be a continuously differentiable function and $J_F(x_n)$ the Jacobian matrix of F. The recursive Newton-Raphson method to find the root of F is given by:

$$\boldsymbol{x}_{n+1} := \boldsymbol{x}_n - \left(J(\boldsymbol{x}_n)^{-1} F(\boldsymbol{x}_n)\right)$$

with an initial guess x_0.

For $f : \mathbb{R} \to \mathbb{R}$ the process is repeated as

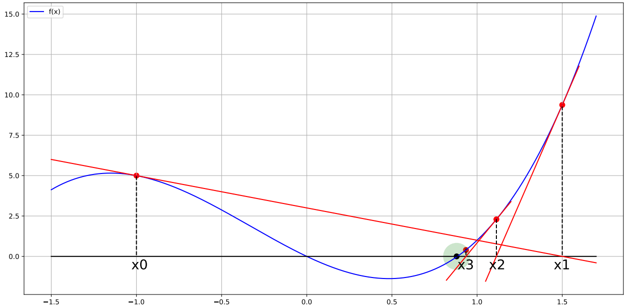
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Accordingly, we can determine the *optimum* of the function f by applying the method instead to $f' = df/dx$.

As an illustrative application, we consider the function

$$f(x) = 3x^3 + 3x^2 - 5x, \quad x \in \mathbb{R},$$

which is represented by the blue line in the following diagram. The figure depicts the iterative solution path applying the Newton-Raphson method to find the root, e.g., x solving $f(x) = 0$, by tangent points and tangents starting from the intial guess $x_0 = -1$.



The first step involves the definition of the function $f(x)$ and its derivation $f'(x)$ in Python. We also specify a delta function that determines the absolute deviation of the target function and the target value, i.e., 0:

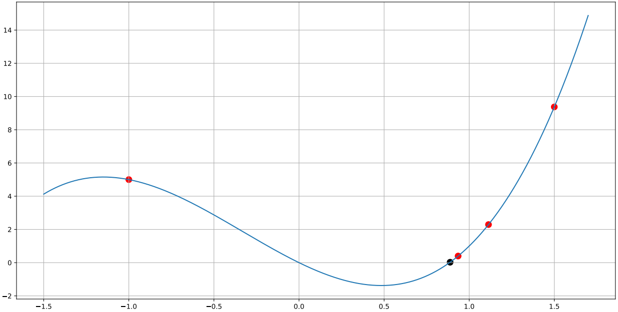
Newton-Raphson requirements

```
def f(x):
    return 3*x**3 + 3*x**2 - 5*x
def df(x):
    return 9*x**2 + 6*x - 5
def dx(f, x):
    return abs(f(x))
```

Finally, we implement the Newton-Raphson algorithm as outlined above. In addition, for a better understanding, we plot the solution path using the tangent points for $x_0, x_1, \ldots, x_N$. The solution point is colored black. Hence, the lines starting with `ax.scatter()` are not part of the algorithm – they take global variables and are included just for the visual illustration.

Newton-Raphson

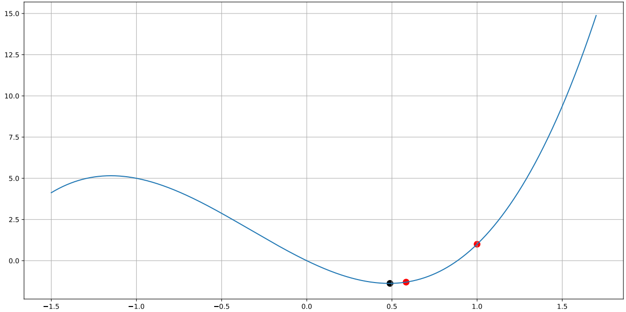
```
def newton_raphson(fun, dfun, x0, e):
    delta = dx(fun, x0)
    while delta > e:
        ax.scatter(x0, f(x0), color="red", s=80)
        x0 = x0 - fun(x0) / dfun(x0)
        delta = dx(fun, x0)
    ax.scatter(x0, f(x0), color="black", s=80)
    return(x0)
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
x = np.arange(-1.5, 1.7, 0.001)
ax.plot(x, f(x))
ax.grid()
x_root = newton_raphson(f, df, -1, 0.1)
fig.savefig("out/newton_raphson_root.pdf")
print(f"Root at: {x_root:.4f}")

## Root at: 0.8878
```


14
12
10
8
6
4
2
0
−2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

With the definition of the second derivative f'', i.e. the derivative of the derivative, we can employ the Newton-Raphson method to obtain an optimum of the target function $f(x)$ numerically. Hence, the previous example needs only minimal modifications:

Newton-Raphson

```python
def ddf(x):
    return 18*x + 6
fig = plt.figure(figsize=(16, 8))
ax = fig.add_subplot(1, 1, 1)
x = np.arange(-1.5, 1.7, 0.001)
ax.plot(x, f(x))
ax.grid()
x_opt = newton_raphson(df, ddf, 1, 0.1)
fig.savefig("out/newton_raphson_optimum.pdf")
print(f"Minimum at: {x_opt:.4f}")

## Minimum at: 0.4886
```


15.0
12.5
10.0
7.5
5.0
2.5
0.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

The End... but not finally